Universidade de Brasília

Instituto de Psicologia

Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

ESTUDOS LONGITUDINAIS E TRATAMENTO DE DADOS AUSENTES EM AVALIAÇÕES EDUCACIONAIS

Luís Gustavo do Amaral Vinha

Brasília

Fevereiro de 2016

Luís Gustavo do Amaral Vinha

ESTUDOS LONGITUDINAIS E TRATAMENTO DE DADOS AUSENTES EM AVALIAÇÕES EDUCACIONAIS

Tese elaborada sob orientação do Prof. PhD. Jacob Arie Laros, apresentada ao Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações da Universidade de Brasília, como requisito parcial a obtenção do título de Doutor em Psicologia Social, do Trabalho e das Organizações.

Brasília

Fevereiro de 2016

Universidade de Brasília

Instituto de Psicologia

Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

_____

Prof. PhD. Jacob Arie Laros (*Orientador*)
Universidade de Brasília – UnB

_____

Prof. PhD. José Francisco Soares (*Membro*)
Universidade Federal de Minas Gerais – UFMG

_____

Prof. Dr. Frederico Neves Condé (*Membro*)
Instituto Nacional de Estudos e Pesquisas
Educacionais Anísio Teixeira – Inep

_____

Prof. Dr. Donald Matthew Pianto (*Membro*)
Universidade de Brasília – UnB

_____

Prof. Dr. Fabio Iglesias (*Membro*)
Universidade de Brasília – UnB

_____

Profa. Dra. Elaine Rabelo Neiva (*Membro Suplente*)
Universidade de Brasília – UnB

Brasília, 26 de fevereiro de 2016

Para Rachel e Isabel, com amor.

Agradecimentos

Agradeço a minha esposa Rachel pelo apoio e dedicação nesse período, obrigado por não deixar de acreditar. A minha pequena, Isabel, pela alegria que você traz para nossas vidas!

Aos meus pais Luiz e Maria Odete, pelas lembranças, dedicação e presença em todas as etapas da minha vida. E aos meus irmãos Luís Henrique e Luís Augusto pela amizade e companheirismo.

Ao meu orientador, professor Laros, pela disponibilidade, atenção dispensada, dedicação e profissionalismo. Muito obrigado!

Aos meus colegas de departamento de Estatística da UnB, pelo incentivo durante esses anos, em especial aos professores Donald, George, Juliana, Joanlise, Maria Teresa, Claudete e Ana Maria.

Aos membros da banca pelas contribuições e ao professor Joaquim José Soares Neto pelas sugestões na etapa de qualificação.

Aos professores do PSTO e aos colegas Fabiana, Felipe, Renata, Alexandre, Talita e Gina. A colega Camila pelas horas compartilhadas com nosso orientador, pelo apoio e sugestões.

Agradeço a Secretaria do Estado do Ceará pela disponibilização dos dados utilizados neste projeto.

Sumário

Lista de figuras

Lista de tabelas

Manuscrito 1. Dados ausentes em avaliações educacionais: comparação de métodos de tratamento.

Manuscrito 2. Tratamento de dados ausentes em uma avaliação educacional com dados longitudinais.

Manuscrito 3. Estudos de fatores associados ao desempenho escolar com dados ausentes não ignoráveis.

Resumo geral

O objetivo desta tese foi contribuir para o desenvolvimento da avaliação educacional por meio de: (a) estudo de simulação visando a avaliar o desempenho de alguns métodos de tratamento e o efeito dos dados ausentes nos resultados, (b) comparação de técnicas estatísticas avançadas usadas para tratamento de dados ausentes nos estudos longitudinais, (c) apresentação de um novo método para tratamento de dados ausentes e (d) aplicação do novo método na identificação de fatores associados ao desempenho escolar, tendo como base os dados de uma avaliação longitudinal. Para isso, foram utilizados os dados da avaliação educacional realizada no estado do Ceará. A tese está dividida em três manuscritos. O primeiro apresenta uma introdução à teoria relacionada aos dados ausentes, as metodologias geralmente utilizadas pelos pesquisadores e os possíveis impactos desses dados nos resultados das pesquisas. Por meio de um estudo de simulação, quatro métodos de tratamentos de dados ausentes (imputação pela média, *listwise deletion*, máxima verossimilhança e imputação múltipla) foram comparados. A imputação pela média apresentou o pior desempenho em todos os cenários e os demais métodos apresentaram resultados semelhantes. Um outro resultado do estudo de simulação foi que o uso de variáveis auxiliares na estimação por máxima verossimilhança e na imputação múltipla reduziu o viés das estimativas quando a ausência simulada não é ao acaso. O segundo manuscrito discute a classificação proposta por Rubin com ênfase na ausência de dados em estudos longitudinais. Esse manuscrito apresenta uma nova metodologia para o tratamento de dados ausentes não ao acaso (MNAR) no contexto de avaliações educacionais. Um estudo de simulação comparou os procedimentos *listwise deletion*, imputação múltipla e a metodologia proposta. Tendo como base o modelo de crescimento linear, verificou-se que o procedimento *listwise deletion* superestimou a taxa média de aprendizado. A imputação múltipla e a metodologia proposta geraram maiores estimativas para os coeficientes das variáveis independentes, e ainda identificaram efeitos de interação. Os resultados evidenciaram a importância da escolha da abordagem a ser utilizada no tratamento de dados faltantes. No terceiro manuscrito, a metodologia proposta para tratamento de dados ausentes foi utilizada no estudo de fatores associados ao desempenho escolar. Em uma amostra composta por 8.681 estudantes do ensino médio, 25,7% estava ausente em pelo menos um momento da avaliação. Verificou-se que a ausência estava relacionada às características dos estudantes e ao desempenho escolar avaliado. A taxa média de aprendizado estimada foi de 8,96 pontos, mas essa taxa varia significativamente entre os estudantes. Com a utilização de dados longitudinais e técnicas de tratamento de dados ausentes, os resultados corroboram estudos transversais de fatores associados ao desempenho escolar. Além disso, demonstra que variáveis relacionadas à idade, número de reprovações e período noturno têm efeitos negativos tanto na proficiência inicial, quanto na taxa de aprendizado.

Palavras-chave: desempenho escolar, tratamento de dados ausentes, dados ausentes não ao acaso, modelo de crescimento linear.

General abstract

The main objective of this thesis was to contribute to the development of educational assessment through: (a) simulation study to assess the performance of some treatment methods and the impact of missing data on results, (b) comparison of advanced statistical techniques for missing data treatment in longitudinal studies, (c) introducing a new method for missing data treatment and (d) application of the new method to identify factors associated with academic performance, based on a longitudinal assessment data. For this purpose, the data of an educational assessment carried out in Ceará State was used. The thesis is divided into three manuscripts. The first manuscript presents an introduction of missing data theory, methodologies generally used by researchers and the potential impacts of such data in research results. A simulation study was used to compare four missing data treatments (mean imputation, listwise deletion, maximum likelihood and multiple imputation). Mean imputation had the worst performance in all scenarios while the other three methods showed similar results. Additionally, the use of auxiliary variables with maximum likelihood estimation and multiple imputation reduced estimation bias when the simulated missingness is not at random. The second manuscript presents the classification proposed by Rubin emphasizing missing data in longitudinal studies. This manuscript proposes a new methodology for the treatment of missing not at random data in the educational assessment context. Listwise deletion, multiple imputation and the proposed methodology were compared in a simulation study. The linear growth model was used for data analysis and comparisons. The average learning rate in Mathematics was overestimated when listwise deletion was used. Multiple imputation and the proposed methodology estimated higher impacts of the independent variables and identified interaction effects. The results highlighted the importance of the method that is chosen to deal with missing data, which is directly related to the assumptions about missing data generating mechanism. In the third manuscript, the proposed methodology for missing data treatment was used to identify factors associated with academic performance. The sample was composed by 8,681 high school students, 25.7% of them were absent at least one moment of follow-up. It was found that the missingness was related to students' characteristics and school performance. The linear growth model showed that the annual learning rate was 8.96 points on average, however it varies significantly among students. Using longitudinal data and missing data treatment techniques, the results corroborate those from cross-sectional studies of factors associated with school performance. Moreover, it shows that variables related to age, school repetition and evening classes have negative effects on both initial proficiency and learning rate.

Keywords: educational assessment, school performance, missing data treatment, missing not at random, linear growth models.

Apresentação

Apesar dos avanços observados nas últimas décadas, a educação no Brasil ainda tem muito a melhorar. Pelos resultados da última edição do PISA (do inglês: *Programme for International Student Assessment Program*) observa-se que os jovens brasileiros aos 15 anos apresentam desempenho escolar insatisfatório, a maioria não compreende o texto que lê e não tem os conhecimentos mínimos esperados em matemática e ciências (OECD, 2013). O Brasil reduziu consideravelmente a taxa de analfabetismo e está próximo da universalização do ensino fundamental (UNESCO, 2014), entretanto, apesar de abrangente, a rede de ensino do país não oferece educação de qualidade.

As avaliações educacionais têm papel fundamental na busca pela oferta de educação com qualidade para todos. Essas avaliações têm como objetivo principal a geração de informação, essencial tanto para apresentar um retrato da realidade, quanto para subsidiar ações. Testes de proficiência e questionários contextuais são utilizados para avaliar o desempenho escolar e obter informações sobre estudantes, professores, diretores e ambiente onde o aprendizado acontece.

A disponibilidade de dados relacionados ao ensino no Brasil é grande. O Inep (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) realiza avaliações periódicas fornecendo informações sobre perfil dos alunos, condições dos estabelecimentos de ensino, matrículas, funções docentes, movimento e rendimento escolar em todo o país. Paralelamente às avaliações nacionais, diversos estados e municípios também têm investido na avaliação de seus sistemas de ensino. Porém, esses dados precisam ser analisados e traduzidos em informações que possam ser usadas para orientar ações de melhoria.

Os dados das avaliações educacionais realizadas no Brasil, especialmente do SAEB (Sistema Nacional de Avaliação da Educação Básica), têm sido utilizados pelos pesquisadores nos estudos de eficácia escolar, principalmente como ferramenta para identificação de fatores associados ao desempenho acadêmico, mas também para análise do efeito escola e avaliação de programas (Karino & Laros, 2016; Vinha & Laros, 2016). De forma geral, os resultados desses estudos apontam a influência da condição socioeconômica e da trajetória escolar dos estudantes no desempenho.

A utilização dos dados dessas avaliações deve considerar as limitações impostas por esse tipo de levantamento. A perda de informação no processo de coleta de dados é comum em avaliações educacionais em larga escala (Cheema, 2014). Essa perda pode ser causada pela ausência dos participantes nos momentos de avaliação, problemas nos instrumentos utilizados, participantes que não detêm o conhecimento ou se recusam a responder parte dos itens do questionário, além de falhas na transcrição dos dados. De qualquer forma, a ausência indesejada de informação tem impacto nos resultados da pesquisa e deve ser avaliada para que esses impactos sejam os menores possíveis.

O objetivo principal desta tese foi contribuir para o desenvolvimento da área de avaliação educacional por meio de: (a) estudo de simulação visando a avaliar o desempenho de alguns métodos de tratamento e o efeito dos dados ausentes nos resultados; (b) comparação de técnicas estatísticas avançadas usadas para tratamento de dados ausentes nos estudos longitudinais;(c) apresentação de um novo método para tratamento de dados ausentes; e (d) aplicação do novo método na identificação de fatores associados ao desempenho escolar, tendo como base os dados de uma avaliação longitudinal. Para isso são utilizados os dados do acompanhamento dos estudantes do ensino médio realizado pelo SPAECE (Sistema Permanente de Avaliação da Educação Básica do Ceará) no período de 2009 a 2011.

O projeto está dividido em três manuscritos. O primeiro manuscrito visa a introduzir a teoria relacionada aos dados ausentes e os métodos de tratamento. O objetivo é apresentar os tipos de dados ausentes, os métodos geralmente utilizados pelos pesquisadores e os possíveis impactos nos resultados. Além dos métodos tradicionais, são apresentados os procedimentos baseados na estimação por máxima verossimilhança e na imputação múltipla. O estudo também compara o desempenho de quatro métodos em diversos cenários simulados, utilizando como base os dados do SPAECE.

O segundo manuscrito avança com a apresentação dos conceitos relacionados aos valores ausentes, enfatizando a ocorrência em estudos longitudinais. Tendo como motivação o padrão de ausência observado na avaliação do SPAECE e os resultados de outros estudos relativos ao abandono e evasão escolar no ensino médio, o manuscrito propõe uma nova abordagem para o tratamento e análise de dados com informações ausentes, considerando que essa ausência não é ao acaso. O método proposto é comparado com dois outros métodos e os resultados são discutidos.

Por fim, o terceiro estudo visa a identificar os fatores que influenciam o desempenho e a evolução dos estudantes do ensino médio, tendo como base a avaliação longitudinal realizada pelo SPAECE. Visando a avaliar o tipo de ausência observada no banco de dados, foi realizada uma comparação dos perfis dos estudantes, agrupados segundo o padrão de ausência observado. A abordagem proposta no segundo manuscrito é utilizada nas análises e os resultados são comparados com estudos anteriores. Espera-se, com esse terceiro estudo, contribuir para a identificação de fatores associados ao desempenho no ensino médio.

Referências

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 20*(10), 1-20.

OECD (2013). PISA 2012 results: *What students know and can do – student performance in Mathematics, Reading and Science (Volume I)*, PISA, Paris: OECD Publishing.

UNESCO (2014). *Teaching and learning: Achieving quality for all.* Paris: UNESCO Publishing.

Karino, C. A., & Laros, J. A. (2016). *Estudos brasileiros sobre eficácia escolar: Uma revisão de literatura*. Manuscrito submetido para publicação.

Vinha, L. G. A., & Laros, J. A. (2016). *Avaliações educacionais no Brasil, Chile e Argentina: Uma revisão de literatura*. Manuscrito submetido para publicação.

Manuscrito 1

DADOS AUSENTES EM AVALIAÇÕES EDUCACIONAIS: COMPARAÇÃO DE MÉTODOS DE TRATAMENTO

Resumo

Os dados ausentes são comuns nas avaliações educacionais. Por isso, o uso de métodos adequados torna-se fundamental para reduzir o impacto da perda de parte da informação. O objetivo do presente estudo é comparar o desempenho de quatro métodos de tratamentos de dados ausentes (imputação pela média, *listwise deletion*, máxima verossimilhança e imputação múltipla) tendo como base o uso de modelos de regressão aplicados aos dados da avaliação educacional realizada no estado do Ceará. Foram utilizadas informações de 7.000 estudantes, simulando-se diversos cenários de acordo com o percentual (10%, 30% e 50%) e o mecanismo gerador da ausência (completamente ao acaso, ao acaso e não ao acaso). A imputação pela média apresentou o pior desempenho em todos os cenários e os demais métodos apresentaram resultados semelhantes entre si. Verificou-se ainda que o uso de variáveis auxiliares na estimação por máxima verossimilhança e imputação múltipla reduziu o viés das estimativas quando a ausência simulada não é ao acaso.

Palavras-chave: tratamento de dados ausentes; avaliação educacional; estudo de simulação.

Abstract

Missing data are common in educational assessments. Therefore, using the appropriate methods becomes vital to reduce the impact of the lost information. The aim of the present study is to compare the performance of four missing data treatments (mean imputation, listwise deletion, maximum likelihood and multiple imputation) based on the results of regression models applied to educational assessment data. Using information of 7,000 students, various scenarios were simulated according to the percentage of missing data (10%, 30% and 50%) and the missingness mechanism (completely at random, at random and not at random). Mean imputation had the worst performance in all scenarios, while the other three methods showed similar results. Additionally, the results indicated that the use of auxiliary variables with maximum likelihood estimation and multiple imputation reduced the bias of the estimates when the simulated missingness is not at random.

Keywords: missing data treatment; educational assessment; simulation study.

Introdução

Apesar da literatura em estatística já abordar o tratamento de dados ausentes há décadas, o assunto ainda é um enigma na pesquisa social aplicada. Muitos pesquisadores não utilizam as técnicas adequadas por falta de familiaridade e, em geral, utilizam métodos simples de eliminação ou substituição, que na maioria das situações não são apropriados (Cox, McIntosh, Reason, & Terenzini, 2014; McKnight, McKnight, Sidani, & Figueredo, 2007; Peugh & Enders, 2004). Trata-se, sem dúvida, de um assunto delicado, pois a ausência da informação pode ser causada por diversos fatores, apresentar diferentes padrões e distorcer os resultados da pesquisa, uma vez que a maioria das técnicas estatísticas foi desenvolvida para dados completos.

A preocupação com o tratamento de dados ausentes na pesquisa social pode ser observada em algumas revisões de literatura. Peugh e Enders (2004) revisaram estudos publicados nas áreas de educação e psicologia aplicada com o objetivo de inventariar como os pesquisadores dessas áreas reportavam a ausência de informação e os procedimentos usados nas análises. Esses autores verificaram que a presença e o tratamento de dados ausentes eram raramente reportados. Na maioria dos casos, os valores ausentes tinham que ser identificados pela comparação do tamanho da amostra e o número de graus de liberdade nas diversas análises realizadas. Além disso, em quase todos os trabalhos onde foi identificada a presença de valores ausentes foram utilizados métodos tradicionais de eliminação ou substituição. Segundo os autores, entre os anos de 1999 e 2003, o número de trabalhos onde os valores ausentes foram reportados aumentou, entretanto, técnicas mais sofisticadas baseadas na estimação por máxima verossimilhança ou imputação múltipla foram pouco utilizadas.

Resultados semelhantes foram encontrados por Rousseau, Simon, Bertrand e Hachey (2012). Esses autores apresentam uma revisão rigorosa de artigos publicados no

British Educational Research Journal durante o período de 2003 a 2007. Dos 68 estudos selecionados, eles observaram que mais de um terço não apresentava qualquer informação sobre dados ausentes. Na metade dos estudos em que a presença de valores ausentes foi reportada, o tratamento adotado não estava explícito e, entre os que mencionaram o tratamento, a grande maioria utilizou métodos de eliminação de observações.

Rousseau e colaboradores (2012) discutem também quais seriam os motivos que levam a tamanha variação nos relatos dos estudos analisados. Os autores acreditam que muitos pesquisadores não mencionam os valores ausentes por ignorar a natureza desses dados, como eles influenciam os resultados, quais métodos estão disponíveis para análise e como reportá-los. Eles afirmam ainda que não existe na área de educação um procedimento que possa ser usado pelos pesquisadores como acontece na psicologia com os manuais da APA (*American Psychology Association*), e que a literatura de estatística relacionada ao tema é muito técnica para esse público. Eles ressaltam também que os pesquisadores não mencionam a existência de valores ausentes, ou reportam resumidamente, dado o aumento na complexidade das interpretações e os efeitos negativos nos resultados gerados pela falta de parte da informação.

No Brasil, considerando os estudos em avaliação educacional, verifica-se também que poucos pesquisadores mencionam a presença de valores ausentes nos dados analisados. Pode-se destacar o estudo de Oliveira, Belluzzo e Pazello (2013), que menciona a retirada das observações com dados faltantes e admite que esse procedimento pode comprometer os resultados observados. Em Xerxenevsky (2012) as observações incompletas também são retiradas, mas nesse caso a autora afirma que a retirada das observações não impactaria nos resultados, dado o tipo de dado ausente presente no estudo. Em alguns trabalhos foram utilizadas variáveis indicadoras no ajuste

dos modelos como uma forma de reduzir o impacto das informações ausentes (Macedo, 2004; Rodrigues, Rios-Neto & Pinto, 2011; Soares & Alves, 2003).

O presente estudo visa a contribuir para o melhor entendimento do impacto dos valores ausentes nos resultados de avaliações educacionais. Para isso, além de discutir conceitos relacionados e apresentar alguns métodos de tratamento e análise de dados com informações faltantes, o desempenho de diferentes procedimentos de análise é comparado em diferentes situações, tendo como base dados reais oriundos de uma avaliação educacional de larga escala realizada no Brasil. Espera-se que este trabalho seja útil para o melhor entendimento do problema, mesmo para leitores que não estão familiarizados com análise de dados e conceitos estatísticos.

Tipos de dados ausentes

Algumas classificações relacionadas a dados ausentes são encontradas na literatura. Por exemplo, McKnight e colaboradores (2007) utilizam uma classificação de ausência de acordo com a fonte: casos ausentes, variáveis ausentes ou ocasiões ausentes. Os casos ausentes acontecem quando o participante não fornece qualquer informação e isso ocorre com a ausência do indivíduo selecionado no momento da coleta de dados. Variáveis ausentes são observadas quando o participante não fornece parte da informação requerida, por exemplo, quando ele não responde um ou mais itens do questionário. Ocasiões ausentes são comuns em estudos longitudinais quando o participante não está presente em todos os momentos de coleta de dados.

Os dados ausentes podem também ser classificados de acordo com o padrão de não resposta (Schafer & Graham, 2002). No padrão univariado apenas uma variável apresenta valores ausentes. Por exemplo, em um modelo de regressão a ausência ocorre somente na variável dependente (Figura 1a). O padrão monótono é geralmente observado em estudos longitudinais em decorrência do abandono de participantes ao

longo das avaliações (Figura 1b). Nesse caso, supondo que *Y* seja a característica acompanhada ao longo do tempo, os indivíduos com valor ausente em Y_t também apresentarão valores ausentes em Y_{t+1}, Y_{t+2} e assim por diante, para qualquer *t*. Por fim, como mostrado na Figura 1c, no padrão arbitrário os valores ausentes podem ocorrer em uma ou mais variáveis para qualquer observação.

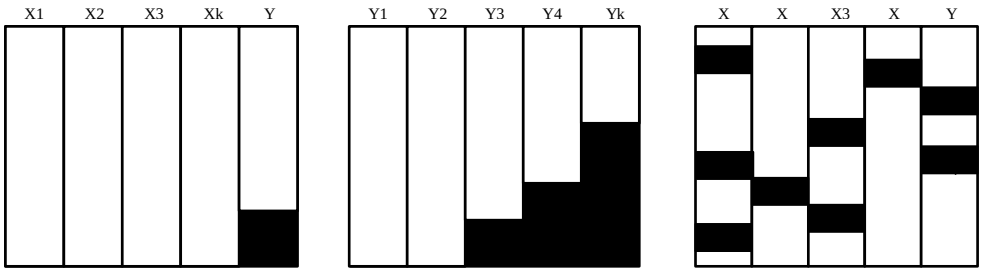

Figura 1
Padrões de não resposta: (a) univariado, (b) monótono, e (c) arbitrário. Adaptado de Schafer e Graham (2002)

Entre as diversas classificações encontradas na literatura a classificação proposta por Rubin é a mais importante (Rubin, 1976; Rubin, 1987). Segundo esse autor os valores ausentes são gerados por mecanismos que relacionam a propensão de ausência aos dados observados. Os dados ausentes podem ser gerados por três mecanismos distintos: valores ausentes completamente ao acaso (MCAR, do inglês *Missing Completely at Random*), ausentes ao acaso (MAR, *Missing at Random*) e ausentes não ao acaso (MNAR, *Missing not at Random*).

Os dados são classificados como ausentes completamente ao acaso (MCAR) quando a ocorrência não está relacionada a qualquer variável observada no estudo ou à própria variável que apresenta os valores faltantes. Dessa forma o mecanismo gerador desses dados não está relacionado a qualquer característica observada, por isso a denominação de ocorrência completamente ao acaso. Esse tipo de ausência poderia ser interpretado como a retirada de uma amostra aleatória de observações do banco de

dados completo. Por exemplo, valores faltantes na proficiência em leitura em uma avaliação educacional pode ser consequência de eventos diversos, como adoecimento do estudante.

Os dados faltantes são classificados como MAR quando a ocorrência está relacionada aos valores observados de outras variáveis, mas independe do valor da variável em questão. Suponha que alunos com baixa renda familiar apresentam maiores taxas de ausência nas provas, no entanto, para uma determinada faixa de renda familiar não é observada qualquer relação entre a ausência e o desempenho (Peugh & Enders, 2004). Nesse caso, o nome dado ao mecanismo pode gerar dúvida (Collins, Schafer & Kam, 2001; Graham, 2009), uma vez que a ausência de dados não é de fato ao acaso. Pode-se então interpretar da seguinte forma: a taxa de valores ausentes está relacionada à renda, mas quando controlada a faixa de renda a ausência passa a ser aleatória (a incidência é aleatória depois do controle de uma variável observada). Na prática não é possível confirmar se a ocorrência de valores ausentes está apenas relacionada com as outras variáveis e não com os valores da própria variável, uma vez que não se conhecem os valores faltantes. No exemplo, entre os indivíduos de baixa renda não é possível verificar se os valores ausentes são dos alunos com menor proficiência (Enders, 2010).

Por fim, os dados faltantes são identificados como MNAR quando a ocorrência está relacionada aos valores da própria variável analisada. Esses dados são mais difíceis de serem identificados, pois a ocorrência está relacionada a valores não observados. Na avaliação educacional isso pode ser exemplificado pela maior taxa de valores faltantes em testes de proficiência entre os alunos com menor desempenho (mesmo depois de controladas outras variáveis).

A classificação proposta por Rubin está diretamente relacionada ao impacto da ausência de informação e à escolha da abordagem mais apropriada para análise dos

dados. Os dados ausentes do tipo MCAR são os que menos influenciam as análises e os resultados, uma vez que a amostra de valores completos pode ser vista como uma amostra representativa da população. Quando os dados faltantes são do tipo MAR a ausência pode ser considerada ignorável, uma vez que se faz necessária a modelagem adicional do mecanismo de ausência de dados no processo de estimação. Por fim, os dados MNAR também são chamados de não-ignoráveis, dado que o mecanismo gerador de ausência deve ser modelado para que sejam obtidas boas estimativas dos parâmetros de interesse (Allison, 2001).

Métodos de tratamento e análise de dados incompletos

Nesta seção, são apresentados métodos tradicionais baseados na eliminação e substituição de valores e dois procedimentos mais sofisticados baseados em métodos de estimação.

De acordo com as revisões de literatura mencionadas anteriormente, os procedimentos baseados na *eliminação de observações* são utilizados na maior parte dos estudos onde os dados ausentes são identificados (Cheema, 2014; Enders & Peugh, 2004; Rousseau & cols, 2012). Acredita-se também que muitos pesquisadores acabam utilizando esses métodos inconscientemente, por ser o procedimento padrão dos pacotes estatísticos (Rousseau & cols, 2012). Os procedimentos baseados na eliminação de observações são *listwise deletion* (ou *complete-case analysis*) e *pairwise deletion* (*available case analysis*).

Listwise deletion

No procedimento *listwise deletion* todos os casos com um ou mais valores ausentes nas variáveis observadas são retirados, logo são consideradas apenas as observações completas. Como consequência, esse procedimento reduz o poder dos

testes estatísticos com a perda de informação e pode gerar estimativas viesadas dos parâmetros quando os dados não são MCAR.

Entretanto, é importante ressaltar que esse método apresenta vantagens que o tornam atrativo em algumas situações. Trata-se de um método simples que não gera maior complexidade nas análises e interpretação de resultados e, como mencionado anteriormente, é o procedimento padrão dos pacotes estatísticos. Em algumas situações o *listwise deletion* apresenta desempenho semelhante a técnicas mais sofisticadas como, por exemplo, quando os dados faltantes são do tipo MCAR, o tamanho da amostra é grande e o número de faltantes é relativamente pequeno (Enders, 2010). Ainda, no ajuste de modelos de regressão, quando os valores ausentes dependem de uma variável presente no modelo (tipo MAR) o procedimento apresenta bom desempenho (Allison, 2001; Schafer & Graham, 2002).

Pairwise deletion

A utilização do procedimento *listwise deletion* gera a perda de uma parcela considerável da informação contida no banco de dados, principalmente quando o número de variáveis envolvidas no estudo aumenta. O procedimento *pairwise deletion* surge então como alternativa para reduzir essa perda de informação.

Nesse caso a parcela de dados utilizada nos cálculos é maior, uma vez que são consideradas observações completas para pares de variáveis e não todo o conjunto de informações de um indivíduo. Esse procedimento é muito comum quando são utilizadas técnicas baseadas em correlações, dado que a estimação da correlação depende apenas de um par de valores. No entanto, essa abordagem é criticada também porque as amostras utilizadas nos diferentes cálculos não são as mesmas, o que pode gerar problemas de estimação e inconsistências de resultados. Ainda, esse procedimento também pode gerar viés se os dados ausentes não são MCAR (Allison, 2001).

As *técnicas de imputação simples* podem ser mais atrativas que os métodos de eliminação, dado que não existe descarte de informação. Todavia, essas técnicas são muito criticadas. Com a utilização de técnicas de imputação simples, os valores imputados são tratados como conhecidos nas análises, o que pode distorcer os resultados. Além disso, alguns pesquisadores evitam a utilização da imputação por acreditar que estariam "construindo os dados" (Graham, 2009). A seguir são apresentados os principais métodos de imputação simples encontradas na literatura.

Imputação pela média

Essa técnica consiste da substituição dos valores ausentes pela média da variável, obtida a partir dos valores válidos observados na amostra. Segundo Enders (2010), trata-se de um procedimento antigo, cuja autoria é frequentemente atribuída a Wilks (1932). A imputação pela média é ainda muito utilizada, mas produz distorções mesmo quando os dados são MCAR. Com a utilização desse procedimento, a variabilidade dos dados pode ser subestimada e, como consequência, há um aumento da chance de se rejeitar hipóteses nulas em testes de significância, além de afetar a estimação de medidas de associação.

O método de imputação pela média é possivelmente o pior tratamento de dados ausentes (Enders, 2010). Utilizando esse procedimento são obtidos estimadores viesados para todos os parâmetros, com exceção da média, para qualquer que seja o tipo de dado ausente analisado.

Imputação pela regressão

A imputação dos dados pela regressão foi proposta por Buck (1960) e, em geral, apresenta melhores resultados que a imputação pela média. Nesse caso, o valor ausente é substituído pelo valor obtido pela equação de regressão utilizando as demais variáveis

que apresentam valores completos como preditoras. Pode ser indicado quando poucas variáveis apresentam observações faltantes. Quando muitas variáveis apresentam dados ausentes esse método perde a atratividade, pois o número de equações a serem estimadas é grande (Enders, 2010).

Como a substituição é feita por um valor predito baseado nas relações entre as variáveis, o valor substituído está exatamente na linha que descreve a relação, o que pode resultar na superestimação das covariâncias entre as variáveis (Enders & Peugh, 2004). Por outro lado, a subestimação da variabilidade é menos acentuada do que a observada com a utilização da imputação pela média e o procedimento gera estimativas não viesadas quando os dados são do tipo MAR.

Para atenuar os possíveis problemas decorrentes desse método, pode-se utilizar substituição pela regressão estocástica (*stochastic regression imputation*). Nesse caso o valor a ser substituído é composto pelo valor predito pela regressão mais uma parcela aleatória, proveniente da distribuição dos resíduos do modelo. Esse método também pode ser encontrado na literatura como *single random imputation* (Cheema, 2014).

Hot deck imputation

O procedimento *hot deck imputation* foi desenvolvido pelo *US Census Bureau* (Enders, 2010). Nesse procedimento, o valor ausente é substituído por um valor observado em unidades similares (Andridge & Little, 2010). Para uma observação com ausência de informação são identificados os possíveis doadores, observações que apresentam valores válidos para a variável a ser tratada e com respostas semelhantes para as demais variáveis coletadas. Por exemplo, o valor observado de apenas um doador selecionado aleatoriamente pode ser usado nessa substituição, ou pode-se utilizar o dado do doador mais próximo, de acordo com alguma métrica pré-estabelecida. Em outras versões do método, os valores referentes a um conjunto de

doadores são combinados, por exemplo, pela média aritmética, e o resultado é então usado na substituição.

Uma característica importante desse método é que a substituição é feita sem a suposição de modelo. Porém, como Andridge e Little (2010) ressaltam, o procedimento depende da medida usada e das variáveis envolvidas na escolha de doadores. O método apresenta bons resultados quando os dados são do tipo MCAR e o número de observações é elevado.

Variáveis indicadoras

O uso de variáveis indicadoras de ausência no ajuste dos modelos foi proposto por (Cohen & Cohen, 1985). Esse método tem sido empregado pelos pesquisadores da área da educação (Cox & colaboradores, 2014) e também pode ser encontrado na literatura brasileira em avaliação educacional (Macedo, 2004; Rodrigues, Rios-Neto, & Pinto, 2011; Soares & Alves, 2003). Esse procedimento visa reduzir a perda de informação e consiste na imputação dos valores faltantes pela média e a inclusão de variáveis que indicam a ausência de informação. Com essa reparametrização um novo intercepto para categorias de dados ausentes é criado e, sob a hipótese de que os dados são MCAR, as estimativas dos parâmetros de interesse não são afetadas (Rodrigues, Rios-Neto, & Pinto, 2011). Contudo, quando os valores ausentes não são MCAR, o procedimento pode gerar viés nas estimativas dos parâmetros de interesse (Cox & colaboradores, 2014).

As abordagens que utilizam métodos de estimação para o tratamento de dados incompletos são apresentadas a seguir. Segundo Enders e Peugh (2004), diversos estudos têm apontado a superioridade desses métodos em relação aos tradicionais apresentados anteriormente. Ainda, Graham (2009) afirma que os pesquisadores deveriam utilizar os procedimentos baseados na máxima verossimilhança e a imputação

múltipla, pois são os melhores métodos disponíveis e são baseados em conceitos fortes e tradicionais da estatística.

Máxima verossimilhança

O tratamento dos valores ausentes pelo método da máxima verossimilhança é semelhante ao utilizado para dados completos (Enders, 2010). O método consiste na escolha de valores para os parâmetros do modelo que maximizam a função de verossimilhança, função que expressa a probabilidade de se observar os valores obtidos na amostra para um determinado modelo escolhido. Esse procedimento requer a suposição de que a ausência é do tipo MAR, além de uma suposição relacionada à distribuição. Em geral assume-se que os dados têm distribuição normal multivariada.

A estimação dos parâmetros é realizada através da maximização da função de verossimilhança, o que na maior parte das situações não pode ser realizado analiticamente. Faz-se necessária a utilização de métodos numéricos, como por exemplo, o algoritmo EM. Esse método de maximização é muito popular quando os dados utilizados na estimação não estão completos (Allison, 2001).

O algoritmo EM é um procedimento iterativo que consiste da repetição de dois passos: estimação (E) e maximização (M). O processo inicia com uma estimação do vetor de médias e da matriz de covariâncias utilizando apenas os dados completos. No passo E os elementos do vetor de médias e da matriz de covariâncias são utilizados para construir um conjunto de equações de regressão, que são usadas para estimar os valores ausentes com base nas variáveis observadas. Em seguida, no passo M, o vetor de médias e a matriz de covariâncias são re-estimados utilizando os estimadores de máxima verossimilhança com base em dados completos (valores observados e os substuídos no passo anterior). Em uma nova etapa E as equações de regressão são re-estimadas e os

valores ausentes são substituídos por novos valores. Na etapa M seguinte, esse novo banco de dados é usado para re-estimar o vetor de médias e a matriz de covariâncias e o processo é repetido até que as estimativas de médias e variâncias não mudem mais (usando um critério de convergência pré-estabelecido).

O algoritmo EM apresenta algumas vantagens importantes, que o tornam muito atrativo (Allison, 2001). Trata-se de um procedimento de fácil implementação, não requer a definição de muitos parâmetros e está disponível em diversos pacotes estatísticos. Entretanto, considerando o ajuste de modelos de regressão, os erros padrão associados aos coeficientes não são obtidos diretamente (Enders, 2010). Como alternativa, pode-se utilizar o procedimento FIML (*full information maximum likelihood*) que, assim como o algoritmo EM, gera estimativas não viesadas para os coeficientes da regressão e ainda estima diretamente os erros associados (Enders, 2001a).

A estimação pelo método da máxima verossimilhança possibilita a utilização de variáveis auxiliares. Essas variáveis não fazem parte do modelo principal de análise, mas estão associadas ao mecanismo gerador de valores ausentes ou a variável a ser tratada. A inclusão de variáveis auxiliares aumenta a chance de satisfazer a suposição de ausência do tipo MAR, e assim melhorar a estimação (Baraldi & Enders, 2010; Collins, Schafer & Kam, 2001). É importante ressaltar que a inclusão de variáveis auxiliares de forma adequada não altera a interpretação dos parâmetros do modelo principal.

De forma geral, a estimação por máxima verossimilhança apresenta diversas propriedades interessantes no tratamento de dados ausentes, entretanto, não se trata de um método infalível. Entre os problemas relacionados, pode-se destacar que a suposição de distribuição normal multivariada geralmente não está satisfeita, o que pode gerar viés nas estimativas e erros associados. Nas situações em que a ausência não é ao acaso

(MNAR) esse procedimento pode distorcer os resultados, apesar de apresentar desempenho superior aos métodos tradicionais, principalmente quando são utilizadas variáveis auxiliares (Graham, 2009).

Imputação Múltipla

Na imputação múltipla, proposta por Rubin (1987), os valores ausentes são repetidamente substituídos por valores obtidos através da simulação da distribuição condicional de probabilidade, tendo como resultado múltiplas versões do banco de dados. Cada versão do banco de dados é analisada de acordo com as técnicas usuais e os resultados são combinados gerando estimativas pontuais dos parâmetros de interesse (Rose & Fraser, 2008).

Esse procedimento é composto por três etapas: imputação, análise e combinação. Na etapa de imputação são gerados *m* novos bancos de dados, essa etapa é composta por dois passos. No primeiro passo (passo I), o vetor de médias e a matriz de covariâncias são estimados e é construído um sistema de equações de regressão para imputação dos valores ausentes, como no método de substituição pela regressão estocástica. No segundo passo (passo P), o vetor de média e matriz de covariâncias são estimados novamente, e a partir dessas estimativas são geradas novas estimativas com a adição de um termo aleatório (esses novos valores correspondem a retirada de uma amostra da distribuição a posteriori da matriz de covariâncias e do vetor de médias). Essas novas estimativas de médias e covariâncias são usadas no passo I seguinte, e o processo se repete até que os *m* conjuntos de dados completos sejam criados.

Algumas considerações importantes são necessárias nessa etapa. Primeiro, a decisão de quais variáveis serão incluídas na etapa de imputação é um aspecto importante para o bom desempenho da imputação múltipla (Enders, 2010). As variáveis presentes no modelo principal devem sempre ser incluídas. Variáveis auxiliares podem

ser incluídas nessa etapa, sem correr o risco de introduzir viés nos resultados, cuidando apenas para não introduzir um número muito elevado de variáveis, o que poderia causar problemas de convergência. Segundo, deve-se considerar um número de iterações antes da retirada do primeiro banco de dados e entre as retiradas dos outros bancos de dados. As iterações iniciais são necessárias para estabilização da distribuição dos parâmetros e as iterações entre as retiradas asseguram a independência entre os dados gerados. Por fim, a eficiência do procedimento está relacionada com o número de bancos de dados gerados (*m*), de forma geral quanto maior o percentual de valores ausentes, maior deve ser *m* (para mais detalhes ver Graham, Olchowski, & Gilreath, 2007).

A etapa de análise consiste na aplicação das técnicas estatísticas usuais aos *m* conjuntos de dados gerados na etapa anterior, de acordo com os objetivos da pesquisa. Como resultado dessa etapa, *m* conjuntos de estimativas para os parâmetros de interesse são gerados. Por fim, na última etapa são calculadas as estimativas dos parâmetros de interesse a partir da combinação das *m* estimativas. A média aritmética das *m* estimativas pode ser usada para gerar a estimativa combinada de um parâmetro, porém essa combinação será válida quando a distribuição do estimador se aproxima da normal. Alguns estimadores têm distribuições assimétricas, como os estimadores de variância e covariâncias, nesses casos transformações podem ser utilizadas para melhorar as estimativas (Enders, 2010).

Os procedimentos de imputação múltipla e máxima verossimilhança são baseados nas mesmas suposições de distribuição normal multivariada e mecanismo gerador de dados ausentes do tipo MAR. Em geral, os resultados obtidos por meio desses métodos são semelhantes, especialmente quando o tamanho da amostra é grande. Segundo Collins, Schafer e Kam (2001), os procedimentos apresentam resultados muito

próximos quando é usado o mesmo banco de dados, com as mesmas suposições de relações entre as variáveis e suas distribuições.

Estudo comparativo

O estudo apresentado neste manuscrito visa a comparar métodos de tratamento de dados ausentes, tendo como base um conjunto de dados reais de uma avaliação educacional. Foram selecionados para essa comparação dois procedimentos tradicionais e que ainda são muito utilizados pelos pesquisadores (Peugh & Enders, 2004): a imputação simples pela média (Me) e o *Listwise Deletion* (LD). E os procedimentos baseados na estimação por máxima verossimilhança (MV) e na imputação múltipla (IM). Além da comparação dos métodos, foi avaliada a utilização de variáveis auxiliares nos procedimentos MV e IM. Este estudo baseia-se na análise dos dados por meio de modelos de regressão linear múltipla, tendo como variável resposta o desempenho escolar.

Dados

Os dados utilizados são provenientes do Sistema Permanente de Avaliação da Educação Básica (SPAECE), disponibilizados pela Secretaria de Educação do Estado do Ceará. Entre outros levantamentos, esse sistema monitora os alunos do ensino médio da rede pública cearense, tendo como medidas de desempenho escolar as proficiências em língua portuguesa e matemática. As proficiências são estimadas por meio da teoria da resposta ao item e expressas em uma escala com média 250 e desvio padrão 50, a mesma utilizada no SAEB (Sistema Nacional de Avaliação da Educação Básica) (Ceará, 2011). Além dos testes os alunos respondem a questionários contextuais com itens relacionados a dados socioeconômicos, hábitos de estudo e clima em sala de aula (http://www.spaece.caedufjf.net/o-programa/).

O banco de dados original disponibilizado contém informações relativas a centenas de milhares de estudantes, com grande incidência de valores ausentes. Entretanto, para a simulação realizada neste estudo foi utilizada uma amostra composta apenas por estudantes com dados completos para as variáveis utilizadas (Tabela 1). Essa amostra contém informação referente a 7.000 estudantes matriculados no 1º ano do ensino médio em 2009 e que estavam presentes na avaliação em 2010. Esses estudantes selecionados foram avaliados por meio das provas de proficiência nos dois anos e responderam ao questionário no primeiro ano. As variáveis utilizadas no estudo são apresentadas na Tabela 1.

Tabela 1
Variáveis utilizadas no estudo

Variáveis	**Descrição / Codificação**
MAT10	Desempenho em Matemática em 2010 (no 2º ano do ensino médio).
MAT09	Desempenho em Matemática em 2009 (no 1º ano do ensino médio).
LP09	Desempenho em Língua Portuguesa em 2009 (no 1º ano do ensino médio).
MASC	0: feminino; 1: masculino.
SUP	Pretensão de ingresso no ensino superior. Obtida a partir da questão relativa aos planos dos alunos após a conclusão do ensino médio. Assume o valor 1 se o aluno pretende ingressar no ensino superior e 0 se tem outros planos.
REPETÊNCIA	Números de vezes que o aluno repetiu um ano escolar, avaliado em 2009. 0: nunca repetiu; 1: uma repetência; 2: duas repetências; 3: três ou mais repetências.
IDADE	Idade reportada em 2009 (em anos).
MANHÃ	Turno em que o aluno frequenta as aulas. Assume o valor 1 se o aluno estudava de manhã em 2009 e 0 se estudava a tarde ou a noite.

Entre os 7.000 estudantes selecionados, 45,3% são do sexo masculino e 43,8% pretendiam ingressar no ensino superior. A idade média observada foi de 15,7 anos (*DP* = 1,1). 32,6% dos estudantes reportaram uma ou mais repetências e 42,0% estudavam no período matutino em 2009. As notas médias em Matemática são 246,5 e 257,1

pontos em 2009 e 2010, respectivamente, com desvios padrão de 46,1 e 46,9 pontos. A nota média em Língua Portuguesa em 2009 foi de 247,5 pontos (*DP* = 39,9).

Procedimentos

A comparação dos métodos foi realizada supondo que essa avaliação seja usada para comparar o desempenho em matemática no segundo ano do ensino médio, segundo o sexo e a intenção de ingresso no ensino superior. Dessa forma, o modelo principal de análise é dado por

$$\mathrm{MAT10} = \beta_0 + \beta_1 \mathrm{MAT09} + \beta_2 \mathrm{MASC} + \beta_3 \mathrm{SUP} + \varepsilon. \qquad (1)$$

A variável MAT09 foi incorporada ao modelo como uma variável de controle para que sejam consideradas as diferenças entre os alunos por meio de uma medida inicial de desempenho. Dessa forma, os coeficientes β_2 e β_3 representam as diferenças entre as proficiências médias dos grupos, controladas pela proficiência no ano anterior.

A Figura 2 apresenta o plano utilizado para a simulação. Foram gerados nove cenários diferentes de acordo com o percentual (10%, 30% e 50%) e tipo de ausência (MCAR, MAR e MNAR). Os dados ausentes foram simulados apenas na variável resposta (MAT10), ou seja, padrão univariado. Para cada cenário foram geradas 100 amostras e cada amostra foi analisada por meio dos quatro procedimentos (Me, LD, MV e IM). Os procedimentos MV e IM foram implementados com e sem variáveis auxiliares para os cenários com dados ausentes do tipo MNAR. Usando este procedimento foram obtidas as estimativas médias dos coeficientes da regressão referentes aos quatro procedimentos em cada cenário.

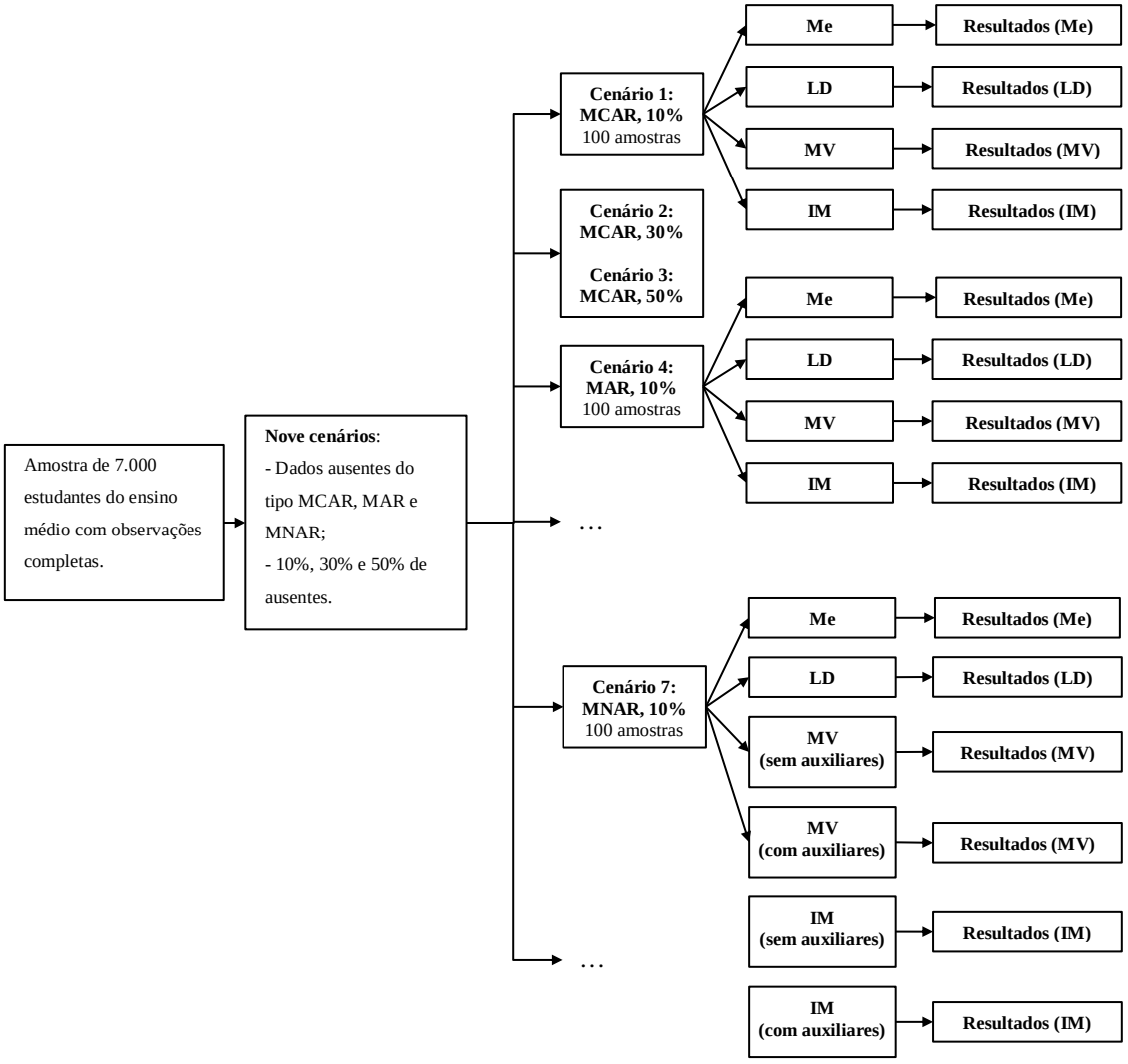

Figura 2
Plano do estudo de simulação

O primeiro mecanismo utilizado gera valores ausentes completamente ao acaso (MCAR). Foram sorteadas amostras aleatórias de indivíduos com valores completos, para esses indivíduos o valor da variável resposta passa a ser ausente. No segundo mecanismo os dados ausentes simulados são do tipo MAR onde a probabilidade de ausência estava relacionada à variável MAT09, de tal forma que a taxa de ausência é maior entre os alunos com menor proficiência em matemática no primeiro ano do ensino médio. Por fim, no terceiro mecanismo foram simulados valores ausentes (MNAR) cuja taxa depende de outras variáveis não presentes no modelo apresentado em (1), nesse caso a ausência simulada está relacionada com as variáveis LP09, IDADE,

REPETÊNCIA e MANHÃ (veja Tabela 1). Nesse último cenário o percentual de valores ausentes é maior entre os alunos com menor proficiência em língua portuguesa, mais velhos, com maior número de repetência e que estudam no período vespertino ou noturno.

O terceiro mecanismo descrito acima pode ser considerado MNAR dado que existe uma dependência residual entre a ausência e a variável que apresenta os valores ausentes (MAT10). Essa dependência se deve a relação existente entre LP09, IDADE, REPETÊNCIA e MANHÃ e a variável MAT10, considerando as demais variáveis do modelo. Essas variáveis foram escolhidas uma vez que estão significativamente relacionadas com a probabilidade de ausência dos alunos no segundo ano do ensino médio (Vinha, 2016[1]).

Análise de dados

As análises foram realizadas utilizando o software estatístico SAS (*Statistical Analysis System*), versão 9.4. Algumas funções específicas foram usadas:

- Para a análise dos dados completos e para o procedimento LD foi usada apenas a PROC REG. Essa função tem como padrão o procedimento *listwise deletion* no ajuste do modelo para dados incompletos. Os modelos foram estimados pelo método dos mínimos quadrados ordinários;
- A estimação por MV foi realizada através do algoritmo EM. A inclusão das variáveis auxiliares (LP09, IDADE, REPETÊNCIA e MANHÃ) foi feita pelo método de estimação em dois estágios[2], para isso utilizou-se a opção EM da PROC MI e a PROC REG;

[1] Manuscrito 3 apresentado neste documento.
[2] Esse método foi proposto por Savalei e Bentler (2009). No primeiro estágio são incorporadas todas as variáveis (de interesse no estudo e auxiliares) na estimação do vetor de médias e matriz de covariâncias pela máxima verossimilhança e no segundo estágio o modelo principal é ajustado considerando somente as variáveis de interesse no estudo.

- Para o procedimento IM foi utilizada a PROC MI com os seguintes parâmetros: m = 50 para os cenários com 10% e 30% de valores ausentes e m = 100 para os cenários com 50% de ausentes, 500 iterações iniciais antes da retirada da primeira amostra e 200 iterações entre as amostras. A PROC MIANALYZE foi usada para gerar as estimativas finais a partir das m amostras.

Resultados e Discussão

As Tabelas 2, 3 e 4 apresentam os coeficientes estimados a partir dos dados completos e as estimativas médias para os diferentes cenários. Essas tabelas mostram também os desvios observados entre as estimativas médias e as estimativas obtidas com os dados completos, de acordo com a equação

$$Desvio\ (\%) = \left(\frac{b_i - b_c}{b_c}\right) \times 100\%,$$

onde b_i e b_c correspondem às estimativas obtidas para os dados incompletos e completos, respectivamente.

Considerando os resultados da amostra completa, pode-se observar que quanto maior a proficiência em matemática em 2009, maior tende a ser a proficiência em 2010, o coeficiente estimado é de 0,677 pontos. Verifica-se que, controlado o desempenho no ano anterior, os meninos têm nota média 3,017 pontos acima da nota média das meninas, e os que pretendem ingressar no ensino superior têm 6,995 pontos a mais, em média. Os resultados apresentados a seguir mostram as variações em relação a esses valores. As variáveis independentes utilizadas no modelo não estão centralizadas, logo o intercepto não pode ser interpretado como a proficiência média dos alunos em 2010.

Primeiramente, pelas Tabelas 2, 3 e 4, pode-se observar que a imputação simples pela média apresenta o pior desempenho, independente do mecanismo gerador e percentual de ausência. Mesmo quando a ausência é completamente ao acaso, as estimativas obtidas por esse método são distantes dos valores reais (Tabela 2). Esses

resultados confirmam que esse procedimento, apesar de simples e de fácil implementação, não deve ser utilizado por apresentar desempenho inferior na maioria das circunstâncias (McKnight & cols, 2007).

Na Tabela 2 são apresentados os demais resultados para os dados ausentes do tipo MCAR. Com exceção da imputação pela média, os resultados observados são semelhantes, as estimativas dos parâmetros do modelo de regressão apresentam pequena variação em relação ao valor observado no ajuste para os dados completos, com desvios de 4,1% ou menos. Verifica-se também que, mesmo com percentual elevado de valores ausentes, a variação observada nos coeficientes ainda é pequena. Em geral, o coeficiente relativo à variável MASC apresenta maiores desvios.

Tabela 2
Resultados para dados ausentes MCAR. Coeficientes do modelo de regressão, com desvio percentual entre parênteses

	Dados completos	Média	LD	MV	IM
10% ausentes					
Intercepto	85,85	**102,8 (19,7)**	85,9 (0,1)	85,9 (0,1)	85,9 (0,1)
MAT09	0,677	0,61 (-9,9)	0,68 (0,4)	0,68 (0,4)	0,68 (0,4)
MASC	3,017	**2,71 (-10,2)**	3,01 (-0,2)	3,02 (0,1)	3,00 (-0,6)
SUP	6,995	6,31 (-9,8)	7,02 (0,4)	7,02 (0,4)	7,01 (0,2)
30% ausentes					
Intercepto	85,85	**136,5 (59,0)**	85,9 (0,1)	85,9 (0,1)	85,7 (-0,2)
MAT09	0,677	**0,48 (-29,1)**	0,68 (0,4)	0,68 (0,4)	0,68 (0,4)
MASC	3,017	**2,08 (-31,1)**	2,99 (-0,9)	2,97 (-1,6)	2,94 (-2,6)
SUP	6,995	**4,95 (-29,2)**	7,00 (0,1)	7,01 (0,2)	7,03 (0,5)
50% ausentes					
Intercepto	85,85	**170,5 (98,6)**	85,89 (0,0)	85,9 (0,1)	85,6 (-0,3)
MAT09	0,677	**0,34 (-49,8)**	0,68 (0,4)	0,68 (0,4)	0,69 (1,9)
MASC	3,017	**1,51 (-50,0)**	3,14 (4,1)	3,14 (3,4)	3,12 (3,4)
SUP	6,995	**3,53 (-49,5)**	7,14 (2,1)	7,13 (1,9)	7,14 (2,1)

Nota: Os valores em negrito correspondem às estimativas com desvio maior que 10%.

O desempenho dos procedimentos LD, MV, IM para análise de dados com ausentes do tipo MCAR deste estudo é semelhante ao observado nos estudos de Enders (2001a) e de Schafer e Graham (2002). Nesses estudos foram utilizados modelos de

regressão aplicados a dados simulados. Enders (2001a) relata que os coeficientes estimados para os quatro métodos avaliados (LD, MV, IM e *pairwise deletion*) apresentam pequenas variações em relação aos verdadeiros parâmetros. Schafer e Graham (2002) avaliaram apenas os métodos de MV e IM e também verificaram pequena variação nas estimativas em relação aos parâmetros quando a ausência é completamente ao acaso.

Pela Tabela 3 pode-se observar que os procedimentos LD, MV e IM têm desempenhos semelhantes quando os dados ausentes estão relacionados com a variável MAT09 presente no modelo. Em comparação com os cenários apresentados na Tabela 2, em geral, os dados ausentes do tipo MAR têm maior impacto na estimação dos coeficientes da regressão. Verifica-se que quanto maior o percentual de ausência, maiores são os desvios observados, e os coeficientes mais afetados estão relacionados com o intercepto do modelo e a variável MAT09. Considerando os coeficientes de MASC e SUP, variáveis não relacionadas diretamente com a ausência de dados, os desvios são pequenos (até 2,6%) quando o percentual de ausência é 10% e 30%, e não ultrapassa 12,8% quando a ausência é simulada em metade das amostras.

Tabela 3
Resultados para dados ausentes MAR. Coeficientes do modelo de regressão, com desvio percentual entre parênteses

	Dados completos	Média	LD	MV	IM
10% ausentes					
Intercepto	85,85	**99,0 (15,3)**	82,79 (-3,6)	82,75 (-3,6)	82,92 (-3,4)
MAT09	0,677	0,63 (-6,9)	0,689 (1,8)	0,689 (1,8)	0,688 (1,6)
MASC	3,017	3,14 (4,1)	2,974 (-1,4)	2,987 (-1,0)	2,999 (-0,6)
SUP	6,995	**5,62 (-19,7)**	7,017 (0,3)	7,019 (0,3)	7,034 (0,6)
30% ausentes					
Intercepto	85,85	**127,7 (48,7)**	**75,30 (-12,3)**	**75,17 (-12,4)**	**75,77 (-11,7)**
MAT09	0,677	**0,54 (-20,2)**	0,717 (5,9)	0,717 (5,9)	0,715 (5,6)
MASC	3,017	3,05 (1,1)	2,940 (-2,6)	3,008 (-0,3)	3,039 (0,7)
SUP	6,995	**3,62 (-48,2)**	7,049 (0,8)	7,074 (1,1)	7,081 (1,2)
50% ausentes					
Intercepto	85,85	**168 (95,7)**	**61,67 (-28,2)**	**61,35 (-28,5)**	**62,65 (-27,0)**
MAT09	0,677	**0,39 (-42,4)**	**0,767 (13,3)**	**0,768 (13,4)**	**0,763 (12,7)**
MASC	3,017	**2,48 (-17,8)**	**2,631 (-12,8)**	**2,711 (-10,1)**	2,796 (-7,3)
SUP	6,995	**2,03 (-71,0)**	6,937 (-0,8)	6,934 (-0,9)	6,984 (-0,2)

Nota: Os valores em negrito correspondem às estimativas com desvio maior que 10%.

No estudo de Enders (2001a) também foram simulados cenários com dados ausentes do tipo MAR. De forma geral, as estimativas dos coeficientes do modelo geradas pelos métodos avaliados apresentam desvios relativamente pequenos. O autor afirma que o método MV apresentou melhor desempenho, porém a diferença entre os métodos é significativa apenas na estimação de um dos coeficientes do modelo. No estudo de Schafer e Graham (2002) não foram observados desvios expressivos nas estimativas dos coeficientes da regressão (para os métodos MV e MI).

Os dados ausentes do tipo MNAR simulados neste estudo têm maior impacto nos resultados (Tabela 4). Para os procedimentos LD, MV e IM, o coeficiente da variável MASC é o mais afetado quando a ausência é não ao acaso. Nota-se que, quando não são consideradas variáveis auxiliares, os métodos MV e IM apresentam resultados praticamente iguais aos observados para o método LD.

Tabela 4
Resultados para dados ausentes MNAR. Coeficientes do modelo de regressão, com desvio percentual entre parênteses

	Dados completos	Média	LD	MV		IM	
				Sem auxiliares	Com auxiliares	Sem auxiliares	Com auxiliares
10% ausentes							
Intercepto	85,85	**100,1 (16,6)**	84,3 (-1,8)	84,3 (-1,8)	83,9 (-2,3)	84,4 (-1,7)	83,9 (-2,3)
MAT09	0,677	0,63 (-6,9)	0,68 (0,4)	0,68 (0,4)	0,69 (1,9)	0,68 (0,4)	0,69 (1,9)
MASC	3,017	3,14 (4,1)	3,15 (4,4)	3,15 (4,4)	3,12 (3,4)	3,15 (4,4)	3,12 (3,4)
SUP	6,995	**5,55 (-20,7)**	6,96 (-0,5)	6,96 (-0,5)	6,97 (-0,4)	6,96 (-0,5)	6,97 (-0,4)
30% ausentes							
Intercepto	85,85	**131,4 (53,1)**	79,5 (-7,4)	79,5 (-7,4)	**74,3 (-13,5)**	79,4 (-7,5)	**74,7 (-13,0)**
MAT09	0,677	**0,53 (-21,7)**	0,71 (4,9)	0,71 (4,9)	0,72 (6,4)	0,71 (4,9)	0,72 (6,4)
MASC	3,017	3,00 (-0,6)	**3,98 (31,9)**	**3,98 (31,9)**	**3,51 (16,3)**	**3,99 (32,3)**	**3,57 (18,3)**
SUP	6,995	**3,40 (-51,4)**	**6,09 (-12,9)**	**6,09 (-12,9)**	6,82 (-2,5)	**6,09 (-12,9)**	6,83 (-2,4)
50% ausentes							
Intercepto	85,85	**164,7 (91,8)**	**75,3 (-12,3)**	**75,3 (-12,3)**	**67,2 (-21,7)**	**75,3 (-12,3)**	**67,3 (-21,6)**
MAT09	0,677	**0,40 (-40,9)**	0,73 (7,8)	0,73 (7,8)	**0,75 (10,8)**	0,73 (7,8)	**0,75 (10,8)**
MASC	3,017	**2,28 (-24,4)**	**4,70 (55,9)**	**4,70 (55,8)**	**3,86 (27,9)**	**4,70 (55,8)**	**3,86 (27,9)**
SUP	6,995	**2,62 (-62,5)**	**5,95 (-14,9)**	**5,95 (-14,9)**	7,06 (0,9)	**5,94 (-15,1)**	7,05 (0,8)

Nota: Os valores em negrito correspondem às estimativas com desvio maior que 10%.

O maior impacto dos valores ausentes do tipo MNAR também é reportado por outros autores. No estudo de Schafer e Graham (2002) os procedimentos MV e MI apresentam estimativas para os coeficientes da regressão distantes dos valores verdadeiros. Em Enders (2001a), desvios expressivos foram observados em apenas uma variável e nesse caso o procedimento LD apresentou os melhores resultados.

Em Langkamp, Lehman e Lemeshow (2010) foram simulados dados ausentes do tipo MNAR, com percentual de ausência variando de 10 a 40%. Nesse estudo foram utilizados dados completos extraídos de uma pesquisa da área de saúde. Os métodos LD e IM têm desempenhos similares quando o percentual de ausentes é igual a 10%, mas quando o percentual é maior o IM gera estimativas mais próximas dos resultados observados para dados completos.

Uma amostra de dados completos obtidos a partir de dados reais também foi utilizada no estudo de Young e Johnson (2013). Os métodos IM, MV e LD apresentam

viés semelhante na estimação dos coeficientes da regressão quando os dados ausentes são do tipo MNAR, porém os autores ressaltam que os métodos IM e MV tem melhor desempenho na estimação dos erros padrão.

Croninger e Douglas (2005) utilizaram dados simulados baseados em variáveis presentes em uma pesquisa realizada entre ex-alunos de uma universidade. Os dados utilizados nas comparações apresentam padrão arbitrário de ausência, contendo ausentes do tipo MCAR, MAR e MNAR. Os métodos da MV e a IM apresentaram resultados superiores em relação a métodos tradicionais, entre eles o LD, com menores desvios em relação aos coeficientes observados para a amostra completa.

No presente estudo verificou-se que a utilização de variáveis auxiliares é importante para reduzir o impacto dos valores ausentes (Tabela 4). Considerando 30% e 50% de valores ausentes, verifica-se que os desvios observados para o coeficiente da variável MASC são reduzidos aproximadamente pela metade. Por exemplo, para o método MV, o desvio passa de 32,1% para 16,5% quando o percentual de ausentes é de 30% e de 55,9% para 27,8% quando 50% estão ausentes. O impacto da ausência de informação observado na variável SUP é menor, com desvios de 12,9% e 15,0% para os maiores percentuais de ausência, e com a utilização das variáveis auxiliares os desvios passam para 2,5% ou menos.

Os benefícios da utilização de variáveis auxiliares são relatados no estudo de Collins, Schafer e Kam (2001). Esse estudo tinha como objetivo principal avaliar o uso de variáveis auxiliares nos procedimentos MV e IM e para isso foram simulados diversos cenários variando os tipos de dados ausentes e a relação entre a taxa de ausência e as variáveis auxiliares. Em geral, eles identificaram que o uso dessas variáveis é importante quando o percentual de ausentes é maior que 25% e a relação

entre a variável auxiliar (usada no mecanismo gerador de valores ausentes) e a variável que apresenta os valores ausentes é forte.

Conclusões e recomendações

A ausência de informação faz parte de praticamente toda pesquisa quantitativa, portanto deve-se fazer o máximo para reduzir seu impacto e assim evitar conclusões equivocadas a partir dos resultados observados. Segundo Allison (2001) os pesquisadores que pretendem mitigar os riscos associados aos dados ausentes devem escolher a estratégia de análise com cuidado, para isso deve-se considerar as características da ausência de informação. Nesse contexto, o presente estudo tem como objetivo principal contribuir para o melhor entendimento do tratamento e análise de dados incompletos.

Neste trabalho foram apresentadas algumas classificações de dados ausentes encontradas na literatura, em especial a proposta por Rubin (Rubin, 1976; Rubin, 1987). Essa classificação é amplamente utilizada e extremamente útil na escolha do tratamento a ser utilizado, no entanto outras informações devem ser analisadas. O pesquisador deve também considerar o objetivo da pesquisa e a técnica principal de análise de dados, a fonte e o padrão de ausência de informação, o tamanho da amostra, o percentual de ausência e a disponibilidade de variáveis que podem ser usadas como auxiliares.

Dada a diversidade de situações e tipos de dados ausentes, não existe um procedimento único e infalível para a análise de dados. Inúmeros procedimentos têm sido desenvolvidos ao longo dos anos e neste artigo foram apresentados resumidamente alguns dos procedimentos mais encontrados na literatura. Buscou-se aqui uma apresentação simples, logo muitos detalhes foram omitidos. Para os leitores interessados, Enders (2010) e McKnight e colaboradores (2007) são boas referências,

esses autores apresentam conceitos e métodos de forma acessível. Para os que buscam uma leitura mais teórica ver Rubin (1987) e Allison (2001).

Outros procedimentos para tratamento de dados ausentes podem ser encontrados na literatura. Por exemplo, é comum na área social o uso de escalas para mensuração de traços latentes (ansiedade, qualidade de vida, motivação), nesses casos os dados ausentes de um item da escala podem ser substituídos pela média dos demais, sob a suposição de que todos representam medidas válidas do mesmo traço latente (Schafer & Graham, 2002). Nos estudos longitudinais a substituição pode ser feita pelo último valor observado para o mesmo indivíduo (McKnight e cols, 2007). Nas ocasiões em que existem casos ausentes (um ou mais indivíduos sem resposta para todo o questionário) podem-se utilizar procedimentos de reponderação (*reweighting*), onde são calculados novos pesos amostrais para os casos presentes na amostra (Schafer & Graham, 2002). Ainda, quando os dados ausentes são do tipo MNAR, os modelos de seleção e de mistura de padrões são muito úteis (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009).

O estudo de comparação realizado teve como foco o ajuste de modelos de regressão aplicados aos dados de uma avaliação educacional. Em geral, os desvios observados nas estimativas dos coeficientes da regressão quando 10% dos dados da variável resposta estavam ausentes foram pequenos. O impacto da ausência de dados depende do tipo de ausência (menor para MCAR e maior para MNAR) e aumenta quando o percentual de ausência é maior. O procedimento *listwise deletion* apresenta resultados semelhantes aos procedimentos baseados na máxima verossimilhança e imputação múltipla nos cenários simulados, e as variáveis auxiliares foram importantes para redução dos desvios.

Os resultados do estudo comparativo apresentado devem ser interpretados com cautela. Optou-se pela utilização de apenas uma técnica de análise de dados na comparação dos métodos, o que limita a extrapolação dos resultados. Foram utilizados modelos de regressão por ser tratar de uma técnica amplamente utilizada na pesquisa aplicada. Em geral, os sistemas de avaliação educacional coletam diversas informações dos estudantes, escolas, professores e diretores, no entanto poucas variáveis foram utilizadas para simplificar a apresentação dos resultados.

O impacto das informações ausentes e a comparação dos métodos foram feitos utilizando apenas as estimativas médias dos coeficientes, porém outros aspectos poderiam ser considerados. Uma análise rigorosa deveria avaliar não somente o viés dos estimadores, mas também o erro quadrático médio[3] (EQM), estimativas dos erros padrão, cobertura de intervalos de confiança, conclusões de testes de significância e medidas de qualidade de ajuste dos modelos (Collins, Schafer, & Kam, 2001; Enders, 2001b). Além disso, nas comparações apresentadas, os coeficientes estimados a partir dos dados completos foram tratados como parâmetros do modelo proposto. Deve-se considerar que existe incerteza associada a esses coeficientes e que eles podem ser afetados por problemas de especificação do modelo, uma vez que não foram estimados a partir de população e sim de uma amostra.

Como o presente estudo teve como base os dados reais de uma avaliação, com variáveis contínuas e categóricas, a suposição de normalidade multivariada não está satisfeita. No entanto, acredita-se que esse fato não comprometa os resultados obtidos uma vez que as estimativas dos coeficientes da regressão são pouco afetadas pela não normalidade (Allison, 2001).

[3] O EQM representa a média das diferenças entre as estimativas individuais e o parâmetro de interesse, elevadas ao quadrado. Dessa forma essa medida considera além do viés, a variância das estimativas (Collins, Schafer, & Kam, 2001).

Referências

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*(1), 40-64.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*, 5-37.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B, 22*, 302-306.

Ceará (2011). Secretaria da Educação. SPAECE – 2011. *Boletim Pedagógico Matemática - Ensino Médio, 3, 1-22*.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 20*(10), 1-20.

Cohen, J., & Cohen, P. (1985). *Applied multiple regression and correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods, 6*, 330-351.

Cox, B. E., McIntosh, K., Reason, R. D., & Terenzini, P. T. (2014). Working with missing data in higher education research: A primer and real world. *The Review of Higher Education, 37*(3), 377-402.

Croninger, R. G., & Douglas, K. M. (2005). Missing data and institutional research. In P. D. Umbach (Ed.), *Survey research: Emerging issues of technology, policy, and analysis* (pp. 33-49). San Francisco: Wiley Interscience Periodicals.

Enders, C. K. (2001a). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61,* 713-740.

Enders, C. K. (2001b). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*, 352-370.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2009). *Longitudinal data analysis.* Boca Raton: Chapman & Hall.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206-213.

Langkamp, D. L., Lehman, A., & Lemeshow, S. (2010). Techniques for handling missing data in secondary analyses of large surveys. *Academic Pediatrics, 10*(3), 205-210.

Macedo, G. A. (2004). *Fatores associados ao rendimento escolar de alunos da 5ª série (2000) - Uma abordagem longitudinal do valor adicionado e da heterogeneidade*. Dissertação de Mestrado. Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford Press.

Oliveira, P. R., Belluzzo, W., & Pazello, E. T. (2013). The public–private test score gap in Brazil. *Economics of Education Review, 35*, 120-133.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525-556.

Rodrigues, C. G., Rios-Neto, E. L. G., & Pinto, C. C. D. X. (2011). Diferenças intertemporais na média e distribuição do desempenho escolar no Brasil: O papel do

nível socioeconômico, 1997-2005. *Revista Brasileira de Estudos de População, 28*(1), 5-36.

Rose, R. A., & Fraser, M. W. (2008). A simplified framework for using multiple imputation in social work research. *Social Work Research*, *32*(3), 171-178.

Rousseau, M., Simon, M., Bertrand, R., & Hachey, K. (2012). Reporting missing data: A study of selected articles published from 2003-2007. *Quality & Quantity, 46*(5), 1393-1406.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling, 16*, 477-497.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Soares, J. F., & Alves, M. T. G. (2003). Desigualdades raciais no sistema brasileiro de educação básica. *Educação e Pesquisa, 29*(1), 147-165.

Vinha, L. G. A. (2016). *Estudo de fatores associados ao desempenho escolar com dados ausentes não ignoráveis*. Manuscrito em preparação.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics, 3*, 163–195.

Young, R., & Johnson, D. (2013). Methods for handling missing secondary respondent data. *Journal of Marriage and Family, 75,* 221-234.

Xerxenevsky, L. L. (2012). *Programa Mais Educação: Avaliação do impacto da educação integral no desempenho de alunos no Rio Grande do Sul.* Dissertação de Mestrado. Pontifícia Universidade Católica, Porto Alegre, RS, Brasil.

Manuscrito 2

TRATAMENTO DE DADOS AUSENTES EM UMA AVALIAÇÃO EDUCACIONAL COM DADOS LONGITUDINAIS

Resumo

A ausência de dados nas avaliações educacionais está relacionada com o perfil dos estudantes e também com o desempenho escolar que está sendo avaliado. O presente estudo tem como objetivo propor uma nova abordagem para o tratamento de dados ausentes não ao acaso, tendo como base o estudo longitudinal realizado no estado do Ceará. Essa nova abordagem, baseada nos modelos de misturas de padrões, é então comparada com outros procedimentos para o tratamento de valores ausentes: *listwise deletion* e imputação múltipla. A amostra utilizada é composta por 8.681 estudantes do ensino médio avaliados entre 2009 e 2011 (26% estavam ausentes em pelo menos um ano do acompanhamento). O modelo de crescimento linear, tendo como variável dependente o desempenho em matemática, foi utilizado para análise dos dados e comparação dos métodos. Os resultados mostraram que, em comparação com as outras abordagens avaliadas, a estimativa da taxa média de aprendizado dos estudantes é maior quando o procedimento *listwise deletion* é utilizado. Os procedimentos de imputação múltipla e mistura de padrões geraram maiores estimativas para os coeficientes das variáveis relacionadas ao trabalho e a intenção de ingressar no ensino superior. Com esses procedimentos também foi possível identificar um efeito significativo da interação entre trabalho e ano letivo. Os resultados evidenciam a importância da escolha da abordagem a ser utilizada no tratamento de dados faltantes, o que está diretamente relacionado às suposições acerca do mecanismo gerador dos dados ausentes.

Palavras-chave: tratamento de dados ausentes; avaliação educacional; estudos longitudinais; imputação múltipla; modelos de misturas de padrões.

Abstract

Missing data are quite common in applied research, especially in longitudinal surveys. This study presents a comparison of three missing data treatment procedures (listwise deletion, multiple imputation and pattern mixture model) applied to a longitudinal study conducted in the state of Ceará in Brazil. The sample consisted of 8,681 high school students evaluated in the period 2009-11 (26% were missing at least one year of follow-up). The linear growth model, with the students' performance in Mathematics as dependent variable, was used for data analysis and comparisons. The results showed that, in comparison with the other methods of missing data treatment, the estimate of average learning rate in Mathematics is higher when listwise deletion was used. Multiple imputation and pattern mixture procedures estimated higher impacts of the variables related to work and higher education intention. In addition, these procedures identify a significant interaction effect between work and school year. The results highlight the importance of the method that is chosen to deal with missing data, which is directly related to the assumptions about missing data generating mechanism.

Keywords: missing data treatment; educational assessment; longitudinal studies; multiple imputation; pattern mixture models.

Introdução

A ocorrência de dados ausentes é comum na pesquisa social aplicada (Rousseau, Simon, Bertrand, & Hachey, 2012). A perda de informação pode ocorrer de diversas formas e, em geral, o problema é maior nos estudos longitudinais, uma vez que os participantes podem se ausentar em qualquer momento da coleta de dados (McKnight, McKnight, Sidani, & Figueredo, 2007; Peugh & Enders, 2004). Levando em consideração a disponibilidade de diversos métodos para a análise de dados incompletos, a escolha da abordagem mais adequada para cada situação não é uma tarefa fácil.

Existem inúmeras razões para a perda de informação nas pesquisas de levantamento (McKnight & cols, 2007). Os valores ausentes podem ser ocasionados por falhas no planejamento ou na execução da pesquisa, por exemplo: parte dos indivíduos da população de interesse é excluída do processo de seleção; indivíduos selecionados não têm o conhecimento necessário para responder a alguns itens do questionário; parte dos itens dos questionários abordam temas delicados e podem constranger o participante; ou ainda, falhas na transcrição dos dados para planilhas podem acontecer, como erros de digitação ou de leitura. Fatores relacionados aos participantes também podem gerar a perda de informação, como nos casos onde um indivíduo selecionado não está presente no momento de coleta de dados por estar doente, ou se recusa a responder alguns itens do questionário.

Nos estudos baseados em coletas longitudinais, a perda de informação também pode ocorrer quando o participante não é acompanhado em todos os momentos da avaliação. Esse tipo de ausência, por sua vez, pode ser classificado como abandono ou ausência com padrão intermitente. O abandono ocorre quando o indivíduo deixa de participar da avaliação a partir de um determinado momento, não retornando mais. No

padrão intermitente, o participante se ausenta em um determinado momento da avaliação, mas retorna em momentos seguintes.

Segundo Cheema (2014), a ausência de dados em uma ou mais variáveis de interesse é constante nas pesquisas educacionais. Segundo o autor, os bancos de dados relativos a avaliações educacionais em larga escala nos Estados Unidos, compostos por milhares de observações, raramente estão completos. A ocorrência de dados ausentes em avaliações educacionais em larga escala no Brasil também é comum. Por exemplo, Karino, Vinha e Laros (2015) verificaram que uma parcela considerável da informação foi perdida no levantamento realizado por meio dos questionários contextuais da edição de 2009 do SAEB (Sistema de Avaliação da Educação Básica). Com base nos dados da edição de 2013 do SAEB, pode-se verificar que mais de 20% dos estudantes registrados no 5º e 9º ano do ensino fundamental não têm o escore relativo aos testes de proficiência, e aproximadamente 30% não responderam ao questionário contextual[4].

As avaliações educacionais com base em levantamentos longitudinais também são muito afetadas por outro fator, a evasão escolar. Logo, além da perda de informação referente à ausência dos estudantes nos momentos da avaliação, parcela considerável de estudantes pode abandonar o estudo por ter se evadido da escola. A evasão escolar ainda é elevada no Brasil, principalmente no ensino médio. Segundo Simões (2014), com base nos dados da PNAD (Pesquisa Nacional por Amostras de Domicílios) de 2012, 45% dos jovens brasileiros de 19 anos não tinha terminado o ensino médio e não estava estudando. O estudo de Shirasu (2014) verificou que aproximadamente 40% dos estudantes das escolas públicas de ensino médio do Ceará, que estavam no 1º ano em 2008, abandonaram os estudos antes do final do ciclo.

[4] Os dados estão disponíveis no endereço: *http://portal.inep.gov.br/basica-levantamentos-acessar*.

Os valores ausentes podem ter impactos significativos na análise dos dados (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009; Schafer & Graham, 2002). A utilização de dados incompletos gera necessariamente redução da precisão, qualquer que seja o padrão da informação perdida. Entretanto, mais grave que a perda de precisão é a possível introdução de viés nas estimativas, o que por sua vez pode gerar conclusões enganosas acerca do fenômeno estudado. A escolha da técnica de análise apropriada depende da identificação do tipo de dado ausente presente no levantamento e é fundamental para minimizar os efeitos dessa ausência de informação.

Neste contexto, o objetivo do presente estudo é propor uma nova abordagem para o tratamento de dados ausentes não ao acaso, baseado no modelo de misturas de padrões. Esse novo método é então comparado com o *listwise deletion* (LD) e com a imputação múltipla (IM). A comparação é realizada tendo como base os dados da avaliação longitudinal realizada no estado do Ceará, no período de 2009 a 2011, por meio do SPAECE (Sistema Permanente de Avaliação da Educação Básica). O modelo de crescimento linear é o modelo principal usado nas comparações, tendo o desempenho dos estudantes em matemática como variável dependente. A seguir são apresentados conceitos relacionados aos dados ausentes e a descrição dos métodos utilizados nas comparações.

Tipos de dados ausentes

A teoria mais conhecida e utilizada para classificação e estudo de dados ausentes foi proposta por Rubin (Rubin, 1976; Rubin, 1987). No presente estudo, os mecanismos geradores de dados ausentes segundo essa classificação são apresentados utilizando o contexto de uma avaliação educacional.

Supõe-se que o desempenho acadêmico de n estudantes é monitorado em T anos por meio de testes de proficiência. Para um determinado estudante i, o conjunto de

respostas planejadas (os escores nos testes de proficiência nos T anos) pode ser representado pelo vetor $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iT})'$. A ausência de valores pode ocorrer em parte dos testes inicialmente planejados no estudo, logo o vetor Y_i pode estar incompleto. A ausência ou presença de informação em Y_i é sinalizada em R_i, de tal forma que $R_{ij} = 0$ quando o estudante i não está presente na avaliação do ano j e $R_{ij} = 1$ quando esse estudante é avaliado no ano j. De acordo com a informação contida em R_i, o vetor Y_i pode ser dividido em Y_i^o e Y_i^a que correspondem às respostas presentes e ausentes, respectivamente.

Os mecanismos geradores de dados ausentes segundo a classificação proposta por Rubin podem então ser definidos de acordo com a distribuição dos indicadores de respostas (R_i), dado os valores de Y_i^o, Y_i^a e X_i, onde X_i corresponde à matriz de covariáveis.

MCAR – Ausentes completamente ao acaso

Os valores ausentes são classificados como completamente ao acaso quando a ausência não está relacionada com as variáveis observadas no estudo ou com a variável que apresenta os valores faltantes. A ausência de valores do tipo MCAR, observada no desempenho dos estudantes em um determinado momento t, não teria então relação com o desempenho a ser avaliado, desempenho anterior ou qualquer outra variável presente no estudo. Assim, a probabilidade de ausência, dado os valores observados e não observados, é dada por

$$P(R_i / Y_i^o, Y_i^a, X_i) = P(R_i).$$

Vale ressaltar que a literatura não é unânime na definição desse tipo de ausência na relação com as covariáveis (X_i). Alguns autores consideram que a ausência pode também ser considerada completamente ao acaso quando a incidência depende de um

conjunto variáveis presentes em X_i, mas independente de Y_i^o e Y_i^a (Demirtas & Schafer, 2003; Fitzmaurice & colaboradores, 2009).

Os dados MCAR podem ser interpretados como uma amostra aleatória de estudantes retirada de um banco de dados completo. Logo, os resultados obtidos utilizando apenas os dados observados podem ser extrapolados para a população de interesse do estudo. Esse tipo de ausência não traz problemas adicionais na análise e interpretação de resultados, além da inevitável perda de precisão resultante da redução do tamanho da amostra. No entanto, a suposição de ausência completamente ao acaso parece pouco adequada no contexto de avaliações educacionais. Por exemplo, a ausência na avaliação pode ser causada pela evasão escolar, e a evasão geralmente está associada a condições socioeconômicas e ao desempenho do estudante (Soares, Fernandes, Nóbrega, & Nicolella, 2015).

É possível avaliar empiricamente a suposição de ausência do tipo MCAR. Uma vez que a perda de informação nessa situação acontece ao acaso, os indivíduos com valores ausentes não diferem daqueles com valores válidos, considerando as características avaliadas no estudo. Por exemplo, supondo que a ausência de valores relativos aos testes de proficiência seja completamente ao acaso, não pode existir diferença de perfil socioeconômico entre os estudantes ausentes e os presentes na avaliação. Dessa forma, essa suposição pode ser avaliada por meio de testes estatísticos tendo como hipótese nula a não diferença entre os grupos, ou seja, que os dados ausentes são do tipo MCAR.

Se por um lado a suposição de ausência completamente ao acaso é muito forte e geralmente não se sustenta na prática, esse mecanismo pode ser usado de forma planejada, com o intuito de diminuir os custos do projeto. Por exemplo, em pesquisas longitudinais em larga escala, pode-se propor que em cada aplicação de testes de

proficiência seja considerada apenas uma parcela dos estudantes envolvidos no estudo, de tal forma que a seleção dos indivíduos seja completamente aleatória e independente das características analisadas (Fitzmaurice & cols, 2009). Assim, é possível obter resultados muito próximos aos que seriam observados com uma amostra completa, mas com menor custo.

MAR – Ausentes ao acaso

Os dados ausentes do tipo MAR têm taxa de ocorrência relacionada às variáveis observadas no estudo (as covariáveis e a variável em questão em momentos anteriores), mas independe do valor da variável em questão. Logo, probabilidade de ausência é dada por

$$P(R_i/\ Y_i^o, Y_i^a, X_i) = \ P(R_i/\ Y_i^o, X_i).$$

Assim, as observações faltantes ocorrem com maior frequência em uma parcela da população. Por esse motivo, os dados completos não podem ser tratados como uma amostra aleatória da população em estudo. Médias e variâncias calculadas com base apenas nas observações completas podem gerar estimativas viesadas para os parâmetros populacionais.

Apesar do potencial de distorcer resultados, a suposição de que o mecanismo gerador de ausência é do tipo MAR traz um resultado importante relacionado a distribuição dos dados faltantes. Por exemplo, suponha que a ausência no escore de desempenho depende da renda familiar do estudante, mas para uma determinada faixa de renda a ausência é aleatória. Logo, em uma certa faixa de renda, não existe diferença entre a distribuição dos valores ausentes e a distribuição dos valores observados de desempenho. Por essa razão, a ausência do tipo MAR também é chamada de ignorável, uma vez que a análise pode gerar inferências válidas, quando utilizadas técnicas adequadas (Allison, 2001; Fitzmaurice & cols, 2009). As técnicas baseadas na máxima

verossimilhança e imputação múltipla têm como suposição a ausência do tipo MAR e apresentam bons desempenhos nesses casos.

Apesar de razoável em muitas situações, a suposição de que os dados ausentes são do tipo MAR não parece adequada em outras (Enders, 2011). Por exemplo, estudantes insatisfeitos com a escola têm maior chance de não responder ao questionário sobre satisfação escolar (Jeličić, Phelps, & Lerner, 2009), nesse caso a ausência depende dos valores não observados da variável dependente.

MNAR – Ausentes não ao acaso

Os dados faltantes são identificados como MNAR quando a ocorrência está relacionada aos valores não observados, além de depender dos valores observados e das covariáveis presentes no estudo. Logo, a probabilidade de ausência é dada por $P(R_i/ Y_i^o, Y_i^a, X_i)$. Uma vez que a ausência de dados não pode ser explicada somente pelos valores observados, esse tipo de ausência também é chamado de não ignorável, ou informativo.

Na avaliação educacional isso acontece quando existe uma taxa maior de valores faltantes nos testes de proficiência entre os alunos com menor desempenho (mesmo depois de controladas outras variáveis). Trata-se do padrão de não resposta mais crítico e pode gerar sérias distorções nos resultados.

É importante ressaltar que a suposição de que os valores ausentes observados são do tipo MAR ou MNAR não pode ser avaliada empiricamente. Essa diferenciação depende dos dados não observados, que em geral não são acessíveis. No exemplo, as informações faltantes geradas pela ausência de estudantes no dia da avaliação não podem ser recuperadas. Esse fato representa um ponto crítico na escolha da abordagem a ser utilizada para análise, uma vez que depende de suposições não testáveis.

Métodos para tratamento de dados ausentes

Técnicas tradicionais de tratamento de dados ausentes, como a imputação pela média ou *listwise deletion*, ainda são muito utilizadas por pesquisadores (Jeličić, Phelps, & Lerner, 2009; Peugh & Enders, 2004; Rousseau e cols, 2012). No entanto, esses métodos têm mostrado desempenho insatisfatório, principalmente quando os dados ausentes não são do tipo MCAR (Schafer & Graham, 2002).

O desenvolvimento de métodos mais sofisticados para análise de dados com observações ausentes tem sido intenso nas últimas décadas. Os métodos baseados na estimação por máxima verossimilhança e na imputação múltipla têm recebido atenção especial por parte dos pesquisadores por apresentarem melhor desempenho que as técnicas tradicionais, principalmente quando os dados ausentes são do tipo MAR (Graham, 2009). Ainda, a literatura relacionada aos métodos de tratamento de dados com ausência não ao acaso (MNAR) também é grande, principalmente na área de bioestatística (Enders, 2011).

Neste estudo os resultados do *listwise deletion* (LD) são comparados com os resultados da aplicação da imputação múltipla e de um procedimento baseado na mistura de padrões (classe de modelos desenvolvidos para análise de dados do tipo MNAR). Essas metodologias são descritas a seguir.

Imputação Múltipla

A imputação múltipla baseia-se na substituição dos valores ausentes realizada *m* vezes, gerando assim *m* versões plausíveis do banco de dados completos (Schafer & Graham, 2002). Essas *m* versões são então analisadas utilizando as técnicas convencionais, com isso as estimativas pontuais e erros padrão são gerados a partir da combinação dos *m* resultados. Sob a suposição de que a ausência de dados é do tipo MAR, a imputação múltipla gera estimativas não viesadas para os parâmetros de

interesse, além disso, a incerteza relacionada à ausência de informação também é considerada nas análises.

Esse procedimento é composto por três etapas: imputação, análise e combinação. Na etapa de imputação são gerados *m* bancos de dados com dados imputados, essa etapa é composta por dois passos. No primeiro passo (passo I), o vetor de médias e a matriz de covariâncias são estimados e é construído um sistema de equações de regressão para imputação dos valores ausentes, como no método de substituição pela regressão estocástica (Enders, 2010). No segundo passo (passo P), o vetor de média e a matriz de covariâncias são estimados novamente, e são geradas novas estimativas com a adição de um termo aleatório (esses novos valores correspondem à retirada de uma amostra da distribuição a posteriori da matriz de covariâncias e do vetor de médias). Essas novas estimativas de médias e covariâncias são usadas no passo I seguinte e o processo se repete até que os *m* conjuntos de dados completos sejam criados.

Na etapa de imputação deve-se considerar o uso de variáveis auxiliares. As variáveis auxiliares não fazem parte do modelo principal de análise, mas estão associadas ao mecanismo gerador de valores ausentes ou a variável a ser tratada (Schafer & Graham, 2002). A inclusão dessas variáveis aumenta a chance de satisfazer a suposição de ausência do tipo MAR, e assim melhorar a estimação (Baraldi & Enders, 2010; Collins, Schafer, & Kam, 2001; Pigott, 2001). Além das variáveis auxiliares, as variáveis presentes no modelo principal devem sempre ser incluídas na etapa de imputação.

A etapa de imputação requer a definição de alguns parâmetros: o número de bancos de dados gerados (*m*), o número de iterações antes da retirada do primeiro banco de dados e entre as retiradas dos demais. A eficiência do procedimento está relacionada ao número de bancos de dados gerados, quanto maior o percentual de valores ausentes

maior deve ser *m* (para mais detalhes ver Graham, Olchowski, & Gilreath, 2007). As iterações iniciais são necessárias para estabilização da distribuição dos parâmetros e as iterações entre as retiradas asseguram a independência entre os dados gerados.

A etapa de análise consiste na aplicação das técnicas estatísticas usuais aos *m* conjuntos de dados gerados na etapa anterior, de acordo com os objetivos da pesquisa. Como resultado dessa etapa, *m* conjuntos de estimativas pontuais e erros padrão são gerados. Por fim, na última etapa são calculadas as estimativas e erros padrão combinados a partir das *m* análises.

A combinação dos resultados pode ser feita pelo método proposto por Rubin (1987). Por esse método, as estimativas pontuais combinadas são calculadas pela média aritmética das *m* estimativas obtidas na etapa anterior. Por exemplo, para o parâmetro β, estimativa pontual é dada por

$$\bar{\beta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i.$$

Além das estimativas pontuais dos parâmetros, os erros padrão também devem ser combinados. Nesse caso, Rubin propôs uma combinação que considera uma variação dentro e entre as amostras. A variação dentro das amostras é dada por

$$V_D = \frac{1}{m}\sum_{i=1}^{m}\widehat{SE}_i^2,$$

onde $\widehat{SE}_i^2$ é o erro padrão ao quadrado relativo ao parâmetro β na amostra i. A variação entre as amostras é calculada por

$$V_E = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\beta}_i - \bar{\beta})^2.$$

O erro padrão combinado é dado pela raiz quadrada de V_T, onde

$$V_T = V_D + V_E + V_E/m.$$

A variação total relacionada à estimação dos parâmetros é composta por uma parcela que corresponde à variação intrínseca dos dados (V_D), mais uma parcela que reflete a variação entre as amostras (V_E), resultado das diferentes imputações. Dessa forma, a imputação múltipla introduz uma incerteza relacionada à ausência de dados, o que não acontece quando a imputação simples pela regressão estocástica é utilizada (Allison, 2001).

Além da suposição de ausência do tipo MAR, a imputação múltipla também requer uma suposição relacionada à distribuição dos dados, em geral a distribuição normal multivariada é utilizada. Os resultados observados para esse método são muito semelhantes aos obtidos pelo método da máxima verossimilhança, especialmente quando o tamanho da amostra é grande (Collins, Schafer, & Kam, 2001).

Modelo de mistura de padrões

Os modelos de mistura de padrões (Glynn, Laird, & Rubin, 1986; Little, 1993; Rubin, 1987) foram desenvolvidos para a análise de dados com valores ausentes do tipo MNAR. Esses modelos são baseados na estimação conjunta do modelo principal de análise e da propensão de ausência de dados.

Considere a seguinte distribuição conjunta

$$p(Y_i, R_i|\theta, \phi),$$

onde p corresponde a distribuição de probabilidade, Y_i é a variável resposta para o indivíduo i (o desempenho do estudante i), e R_i é o indicador de valor ausente correspondente. O termo θ corresponde ao conjunto de parâmetros que descrevem a distribuição de Y (por exemplo, por meio de um modelo de crescimento linear), e ϕ contém o conjunto de parâmetros que descrevem a propensão de ausência de dados em Y (através de um modelo de regressão logística). No modelo de mistura de padrões, essa distribuição conjunta é expressa pelo seguinte produto de distribuições

$$p(Y_i, R_i|\theta, \phi) = p(Y_i|R_i, \theta)\,.\,p(R_i|\phi),$$

onde $p(Y_i|R_i, \theta)$ é a distribuição condicional de Y_i, dado um particular valor de R_i, e $p(R_i|\phi)$ a distribuição marginal de R_i. Por essa decomposição o primeiro termo corresponde ao modelo principal de análise dos dados para grupos de indivíduos que compartilham o mesmo padrão de ausência, e o segundo termo corresponde ao modelo que descreve a incidência de valores ausentes desses grupos.

A aplicação dessa classe de modelos consiste na estratificação da amostra em grupos de acordo com o padrão de ausência e, para cada grupo, o modelo principal de análise é ajustado (Enders, 2011). As estimativas finais são obtidas utilizando os resultados de cada grupo, considerando o tamanho dos grupos na ponderação. Contudo, para alguns desses grupos, o modelo não pode ser estimado devido à ausência de valores, portanto para que esse modelo possa ser usado são necessárias suposições adicionais, também chamadas suposições de identificação.

Por exemplo, considere a aplicação de um modelo de crescimento linear em um levantamento com três momentos distintos de avaliação, e dados ausentes causados pelo abandono de parte dos indivíduos. Os grupos formados de acordo com o padrão de ausência são: o grupo composto por indivíduos presentes apenas no primeiro momento de avaliação; o segundo com os que estavam presentes no primeiro e no segundo momento; e o grupo presente em todos os momentos de coleta de dados. Para cada grupo, considerando que os tamanhos das amostras são adequados, o modelo de crescimento deve ser estimado. No entanto, como a variável resposta foi observada apenas em um momento, o modelo proposto não poderia ser estimado para o primeiro grupo. Logo, uma suposição de identificação relacionada a desempenho desse grupo torna-se indispensável.

Uma estratégia para contornar o problema da estimação é a utilização da combinação de padrões de ausência. No exemplo acima, o pesquisador poderia supor que o padrão de desempenho dos indivíduos presentes apenas no primeiro momento é o mesmo padrão observado para os que estavam presentes nos dois primeiros momentos de avaliação. Essa estratégia, proposta por Hedeker e Gibbons (1997), faz com que a amostra seja dividida em dois grupos em que o modelo de crescimento linear possa ser ajustado.

Outra estratégia utilizada para a estimação do modelo de mistura de padrões é a imposição de restrições. Nesse caso, o parâmetro que não pode ser estimado para um determinado grupo é substituído por estimativas obtidas em outros padrões. A imposição de restrição pode ser feita utilizando: casos completos, casos vizinhos e casos disponíveis. Quando são utilizados os casos completos, o parâmetro não identificado no grupo é substituído pela estimativa obtida para indivíduos com dados completos. De forma semelhante, a imposição pode ser feita utilizando os casos vizinhos, ou seja, os grupos com padrão de ausência mais próximo. Ou ainda, utilizando todos os casos disponíveis, a estimativa utilizada na substituição é obtida a partir das estimativas de todos os grupos onde o parâmetro pode ser estimado. Mais detalhes e outras estratégias de identificação são apresentados em Demirtas e Schafer (2003).

Segundo Enders (2011) o modelo de misturas de padrões pode ser mais interessante do que outros presentes na literatura, como por exemplo, os modelos de seleção. O modelo de misturas de padrões requer um tipo de suposição que obriga o pesquisador a explicitar a escolha, por outro lado, nos modelos de seleção, a suposição é imposta a uma distribuição, o que pode ser vago. Além disso, o modelo de mistura de padrões possibilita estudos de sensibilidade através da utilização de diferentes restrições.

Método proposto

Antes de escolher a abordagem a ser utilizado na análise dos dados, o pesquisador deve avaliar as suposições mais adequadas em relação ao mecanismo gerador da ausência (Pigott, 2001). Detalhes sobre o processo de coleta de dados e o conhecimento da área de pesquisa são fundamentais nessa avaliação. Os dados a serem analisados no presente estudo referem-se ao acompanhamento dos estudantes do ensino médio no Estado do Ceará. A trajetória desses alunos é influenciada pelo contexto socioeconômico e cultural, e esse fato deve ser considerado na escolha da abordagem. Alguns resultados importantes relativos à trajetória de alunos do ensino médio no Brasil foram usados para a formulação do procedimento a ser utilizado.

De Leon e Menezes-Filho (2002) analisaram os fatores relacionados à reprovação, avanço e evasão escolar por meio dos dados da Pesquisa Mensal de Emprego (PME), entre 1985 e 1997. Esse estudo indicou que idade, sexo, renda, trabalho e composição familiar são importantes para explicar a trajetória escolar no ensino médio. Com base nos dados da PNAD (Pesquisa Nacional por Amostra de Domicílios) de 2004 e 2006, Neri (2009) afirmou que a falta de interesse pela escola e a necessidade de renda são os fatores mais importantes para explicar a evasão escolar entre os indivíduos de 15 a 17 anos. Soares e colaboradores (2015) identificaram que o abandono escolar no ensino médio em Minas Gerais é afetado pela composição familiar, defasagem idade/série, trabalho, condição socioeconômica, gravidez e dificuldade nas disciplinas.

Utilizando os dados do ensino médio do SPAECE, Shirasu (2014) verificou que a chance de evasão entre os indivíduos com reprovações é o dobro da chance dos indivíduos sem reprovação. Além disso, a autora identificou que as variáveis sexo, distorção idade/série e escolaridade dos pais são importantes preditoras tanto da repetência quanto da evasão escolar. Ainda com base nos dados da avaliação no Ceará,

Vinha (2016b)[5] verificou que a reprovação está fortemente associada a ausência no último ano do acompanhamento do ensino médio.

Por esses resultados, optou-se pela elaboração de procedimento que tenha como suposição a ausência do tipo MNAR. A escolha da classe de mistura de padrões nesse caso é motivada pela forte relação entre a repetência e a evasão escolar, além da relação entre a evasão e outras características dos estudantes, como renda, escolaridade da mãe, trabalho e ao próprio desempenho. Esse procedimento propõe a mistura de padrões entre os repetentes e os ausentes. A suposição principal é que os estudantes com valores ausentes no desempenho teriam evolução semelhante aos estudantes com reprovação no ensino médio. Essa suposição baseia-se na relação entre a reprovação e a evasão, ambas estão relacionadas a características socioeconômicas, mas também são resultantes da falta de motivação e o baixo desempenho dos estudantes nas disciplinas.

Verificou-se ainda que, dentro dos grupos formados pela combinação de padrão de ausência e repetência, ainda existia grande heterogeneidade no perfil dos estudantes. Foi proposta então, para cada grupo, uma estratificação baseada nas variáveis que estão relacionadas com a ausência de dados. Vinha (2016b) verificou que idade, turno, escolaridade da mãe, trabalho e o desempenho em língua portuguesa são importantes preditoras da ausência na variável desempenho. Assim, visando a criar estratos mais homogêneos de alunos, foi proposta uma subdivisão dos grupos de acordo com o escore de propensão a ausência a partir dessas variáveis.

Por fim, o procedimento utiliza a imputação múltipla para a geração das estimativas pontuais e erros padrão. Nesse caso, uma modificação foi proposta no procedimento, a etapa de imputação foi realizada separadamente para cada estrato

[5] Manuscrito 3 apresentado neste documento.

proposto, de tal forma que ocorra a mistura de padrões de acordo com as suposições acima.

Método

Dados

Os dados utilizados referem-se à avaliação dos estudantes do ensino médio realizada pelo SPAECE. Nesse sistema, os estudantes do ensino médio são avaliados nos três anos do ciclo por meio de testes de proficiência em matemática e língua portuguesa e respondem a questionários contextuais. As proficiências são estimadas por meio da teoria da resposta ao item e expressas na escala com média 250 e desvio padrão 50 utilizada no SAEB (Sistema Nacional de Avaliação da Educação Básica) (Ceará, 2011).

Centenas de milhares de estudantes são avaliados anualmente pelo sistema, porém grande parte da informação é perdida. Para o presente estudo foi selecionada uma amostra de alunos que cursavam o primeiro ano do ensino médio em 2009. Além das informações de 2009, foram consideradas as proficiências avaliadas nos dois anos seguintes, 2010 e 2011, o que corresponde ao segundo e terceiro ano para os estudantes sem reprovação, e também a proficiência avaliada em 2008, quando esses estudantes estavam no último ano do ensino fundamental.

Os arquivos de dados referentes a essas avaliações não têm uma identificação comum para os estudantes de um ano para outro. Como consequência, os arquivos de dados não podem ser interligados diretamente. Nesse caso, o nome dos estudantes foi utilizado como indexador. No entanto, esse campo pode apresentar diversos erros, dificultando assim a identificação. Mais detalhes sobre a junção dos dados são apresentados no Apêndice 1.

Após o processo de interligação dos arquivos de dados, uma base composta pelas informações relativas a 8.681 estudantes foi consolidada. Da forma com que a amostra foi selecionada, todos os estudantes estavam presentes na avaliação de 2009, logo todos têm o registro dos testes de proficiência e as informações referentes ao questionário contextual naquele ano. As proficiências relativas a 2010 e 2011 estão incompletas, como apresentado na Tabela 1. Observa-se que 450 estudantes não estavam presentes na avaliação de 2010, mas estavam em 2011, logo esses indivíduos não se evadiram da escola e a ausência foi devida a outro motivo desconhecido. Entre os que estavam presentes somente em 2009, ou em 2009 e 2010, com base nos dados disponibilizados, não é possível identificar se eles apenas faltaram nos dias das avaliações ou se a ausência foi causada por abandono ou evasão escolar[6]. Nesse caso, seria necessária a utilização de outras fontes de informação, como o Censo Escolar, para identificar se os estudantes ausentes estavam matriculados na escola ou não.

Tabela 1
Padrões de ausência de dados

Padrão	2009	2010	2011	Número de estudantes
0	P	P	P	6.447
1	P	A	P	450
2	P	P	A	652
3	P	A	A	1.132
Total				8.681

Nota: P = Presente e A = Ausente.

Além das proficiências, algumas variáveis relacionadas aos alunos foram utilizadas no presente estudo. Essas variáveis foram coletadas por meio do questionário

[6] Considera-se que o aluno está afastado por abandono quando ele deixa de ir para a escola em um determinado ano letivo, mas retorna no(s) ano(s) subsequente(s). A evasão escolar acontece quando um aluno, reprovado ou aprovado em um determinado ano, não se matricula nos demais anos do estudo (Gonçalves, 2008).

respondido pelos estudantes em 2009. Na Tabela 2 são apresentadas as variáveis utilizadas no estudo.

Tabela 2
Variáveis utilizadas no estudo

Variáveis	**Descrição / Codificação**
MAT09	Desempenho em Matemática em 2009.
LP09	Desempenho em Língua Portuguesa em 2009.
MAT10	Desempenho em Matemática em 2010.
MAT11	Desempenho em Matemática em 2011.
MAT08	Desempenho em Matemática em 2008, no último ano do ensino fundamental.
Sexo	0: feminino; 1: masculino.
Etnia	Etnia autodeclarada: branco, pardo, negro, amarelo e indígena.
Turno	Turno em que o estudante frequentava as aulas: manhã, tarde ou noite.
Idade	Idade reportada em 2009 (em anos).
Escolaridade da mãe	A escolaridade da mãe é classificada em: - nunca estudou - 1ª a 4ª série do ensino fundamental - 5ª e a 8ª série do ensino fundamental - 1ª e a 3ª série do ensino médio - ensino superior - não sabe.
Repetência no ensino fundamental	Número de reprovações ensino fundamental: nunca repetiu; 1 reprovação; 2 reprovações; 3: três ou mais repetências.
Superior	Pretensão de ingresso no ensino superior. Obtido a partir da questão relativa aos planos dos alunos após a conclusão do ensino médio. Assume o valor 1 se o aluno pretende ingressar no ensino superior e 0 se tem outros planos.
Gosta de Matemática	Matemática é a disciplina que o estudante mais gosta. Assume o valor 1 se o aluno responde que Matemática é a disciplina preferida e 0 se prefere outra.
Dever	Assume o valor 1 se o aluno faz as tarefas de casa sozinho e 0 se não faz sozinho ou é auxiliado por outra pessoa.
Repetência no ensino médio	Números de vezes que o aluno repetiu um ano escolar no ensino médio. Obtido a partir do registro da etapa cursada em cada ano. 0: nunca repetiu 1: uma repetência 2: duas repetências
Trabalho	O estudante trabalha fora de casa. Obtido a partir da questão relativa ao que o aluno faz quando não está na escola. Assume o valor 1 se o aluno trabalha fora de casa e 0 se tem outras atividades.

Procedimentos

A comparação apresentada neste estudo baseia-se na estimação do modelo de crescimento linear (Raudenbush & Bryk, 2002). Esse modelo é apresentado considerando uma estrutura em dois níveis: o primeiro nível corresponde às observações "dentro" de cada indivíduo e o segundo é o nível dos indivíduos. Dado o número reduzido de observações para cada estudante, o componente temporal utilizado é descrito por uma função do primeiro grau. Nesse caso a equação que descreve o primeiro nível é dada por

$$Y_{ij} = \pi_{0j} + \pi_{1j}(ANO)_i + \varepsilon_{ij},$$

onde Y_{ij} é a proficiência do aluno j, no ano i; π_{0j} é a proficiência esperada do aluno j no início do ensino médio[7]; π_{1j} é a taxa de aprendizado do aluno j em um ano acadêmico; e ε_{ij} é a parcela aleatória da proficiência não explicada pelo modelo.

A proficiência em Matemática no final do ensino fundamental (*MAT08*), a pretensão de ingresso no ensino superior (s*uperior*) e o trabalho fora de casa (*trabalho*) foram utilizadas como variáveis independentes no nível dos alunos. As equações para esse nível são:

$$\pi_{0j} = \beta_{00} + \beta_{01}(Trabalho)_j + \beta_{02}(Superior)_j + \beta_{03}(MAT08)_j + r_{0j} \text{ e}$$

$$\pi_{1j} = \beta_{10} + \beta_{11}(Trabalho)_j + r_{1j}.$$

Por essa formulação, as variáveis *MAT08* e *superior* têm influência na proficiência inicial e a variável *trabalho* é importante tanto para a proficiência inicial quanto para a taxa de crescimento. Nesse caso, o coeficiente β_{01} representa a diferença no desempenho no início do ensino médio entre o grupo de estudantes que trabalham e os que não trabalham. O coeficiente β_{11} quantifica a mudança na taxa anual de

[7] Como as avaliações acontecem no final do ano, optou-se pela parametrização onde a variável ANO assume o valor 1 para 2009, 2 para 2010 e 3 para 2011. As variáveis independentes utilizadas no modelo foram centradas na média para facilitar a interpretação dos resultados.

crescimento para os estudantes que trabalham em relação aos demais. Os termos r_{0j} e r_{1j} correspondem aos componentes aleatórios da proficiência esperada e da taxa de aprendizado.

O procedimento de mistura de padrões proposto neste estudo baseia-se na divisão dos dados para a estimação com dados ausentes. Nesse caso, o banco de dados foi dividido de acordo com o padrão de ausência e o escore de propensão a ausência no final do estudo. Inicialmente os estudantes foram divididos em dois grupos: os estudantes sem reprovação no ensino médio e que apresentam padrão 0 e 1 de ausência (Tabela 1); e os demais, ou seja, todos os estudantes com uma ou duas reprovações no ensino médio ou que apresentavam padrão 2 ou 3 de ausência. Em sequência, uma nova divisão é proposta, esses grupos são então divididos de acordo com o escore de propensão. Para isso foram calculados os decis do escore de propensão dentro de cada grupo e então os estudantes foram divididos em 5 subgrupos tendo como limites o 2º, 4º, 6º e 8º decil. A Figura 1 apresenta esquematicamente essa divisão.

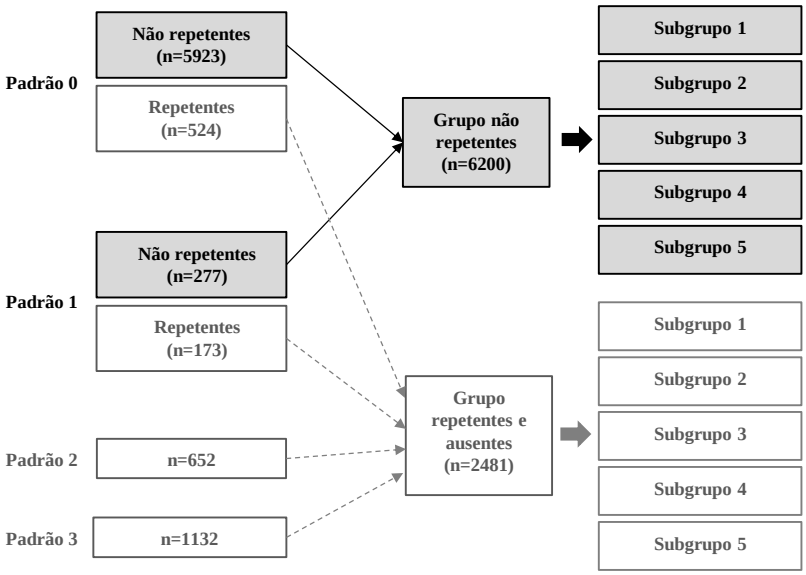

Figura 1
Divisão dos estudantes no procedimento de mistura de padrões

Análise de dados

Todas as análises foram realizadas utilizando o software estatístico SAS (*Statistical Analysis System*), versão 9.4. Para o procedimento *listwise deletion* foi utilizada a PROC MIXED para o ajuste do modelo de crescimento linear proposto. Nesse caso foram utilizadas apenas as observações completas.

No procedimento de imputação múltipla foram utilizadas as funções MI, MIXED e MIANALYZE do SAS. Na etapa de imputação foram utilizadas as variáveis *MAT08, LP09, sexo, turno, idade, escolaridade da mãe, superior, gosta de Matemática, dever, repetência no ensino fundamental e trabalho,* considerando todos os estudantes em uma única etapa de imputação. Nessa etapa foi utilizada a PROC MI com os seguintes parâmetros: *m*=100, 500 iterações iniciais antes da retirada da primeira amostra e 200 iterações entre as amostras. A PROC MIXED foi utilizada na etapa de estimação, o modelo de crescimento linear proposto foi então estimado com base nas 100 amostras geradas na etapa anterior. A PROC MIANALYZE foi usada para gerar as estimativas finais a partir dos resultados das *m* estimativas obtidas na etapa anterior segundo o método proposto por Rubin (1987).

O modelo de mistura de padrões proposto neste estudo para análise dos dados também utiliza as funções MI, MIXED e MIANALYZE do SAS. No entanto, nesse caso a etapa de imputação é realizada separadamente em cada estrato formado pela combinação de grupo, repetência e escore descrita anteriormente. Os escores utilizados foram gerados pela regressão logística (PROC LOGISTIC) tendo como variável resposta a ausência no último ano do estudo e como independentes as variáveis *MAT09, LP09, turno, escolaridade da mãe, idade, sexo, repetência no ensino fundamental, trabalho e superior*.

Resultados

A Tabela 3 apresenta a distribuição das variáveis coletadas por meio dos questionários. Entre os 8.681 estudantes selecionados na amostra, 54,1% eram estudantes do sexo feminino, com idade média de 15,83 anos (*DP=1,06*) e 39,7% estudavam no turno da manhã em 2009. Vale destacar que o registro da repetência no ensino médio foi possível apenas para os estudantes presentes nos anos de 2010 ou 2011 ou em ambos, nesse caso a estimativa da taxa de reprovação no ensino médio foi obtida com base no registro de 7.549 estudantes. Entre esses estudantes, 10,56% repetiram uma ou duas vezes no ensino médio.

Tabela 3
Análise Descritiva

Variável	%	Variável	%
Sexo		*A disciplina preferida é Matemática*	
Feminino	54,12	Não	81,53
Masculino	45,88	Sim	18,47
Turno		*Faz o dever de casa sozinho*	
Manhã	39,73	Não	36,3
Tarde	35,2	Sim	63,7
Noite	25,07		
		Trabalha fora de casa	
Etnia		Não	84,6
Branco	19,38	Sim	15,4
Pardo	58,05		
Negro	12,17	*Número de repetências no EF*	
Amarelo	6,49	0	62,38
Indígena	3,91	1	24,08
		2	10,14
Escolaridade da mãe		3 ou mais	3,41
Nunca estudou ou não sabe	22,16		
1a, a 4a, do EF	31,71	*Número de repetências no EM**	
5a, a 8a, do EF	22,23	0	89,35
1a, a 3a, do EM	17,53	1	9,43
Superior	6,36	2	1,22
Superior			
Não	59,27		
Sim	40,73		

* Os percentuais referem-se ao total de 7.549 estudantes.

As trajetórias dos alunos de acordo com o padrão de ausência e a repetência no ensino médio foram analisadas, considerando apenas os valores válidos da amostra. Pode-se observar, pela Figura 2, que os alunos presentes nas três avaliações realizadas no ensino médio têm proficiência média maior que os demais e apresentam maior taxa de crescimento no período.

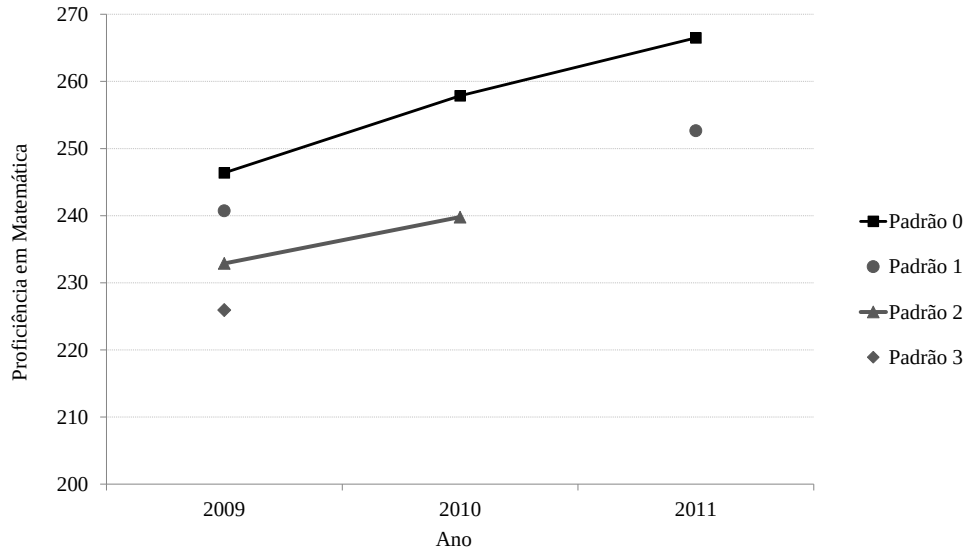

Figura 2
Proficiência média dos estudantes dos diferentes padrões de ausência

O gráfico da Figura 2 sinaliza alguns pontos importantes que foram considerados no tratamento dos dados. Primeiro, os indivíduos sem valores ausentes têm desempenho médio superior em relação aos demais. Segundo, a comparação das trajetórias dos estudantes com algum tipo de ausência (padrões 1, 2 e 3) sugere que a evasão escolar pode estar associada aos piores desempenhos. Por exemplo, a ausência dos estudantes do padrão 3 pode ter sido causada pela evasão, abandono ou outro motivo desconhecido que levou o estudante a não comparecer nos dias das avaliações. Por outro lado, os estudantes do padrão 1 não se evadiram da escola, a ausência em 2010 foi causada pelo abandono naquele ano ou outro motivo desconhecido. Logo, o desempenho inferior dos

estudantes do padrão 3 pode ser atribuído àquela parte de estudantes que se evadiram da escola.

A reprovação no ensino médio também é importante na evolução da proficiência em Matemática dos estudantes no período analisado. Pela Figura 3, verifica-se que os estudantes que repetiram pelo menos um ano escolar durante o ensino médio têm desempenho inferior aos que não repetiram. Pode-se observar que os repetentes têm proficiências médias menores e também uma evolução menos acentuada.

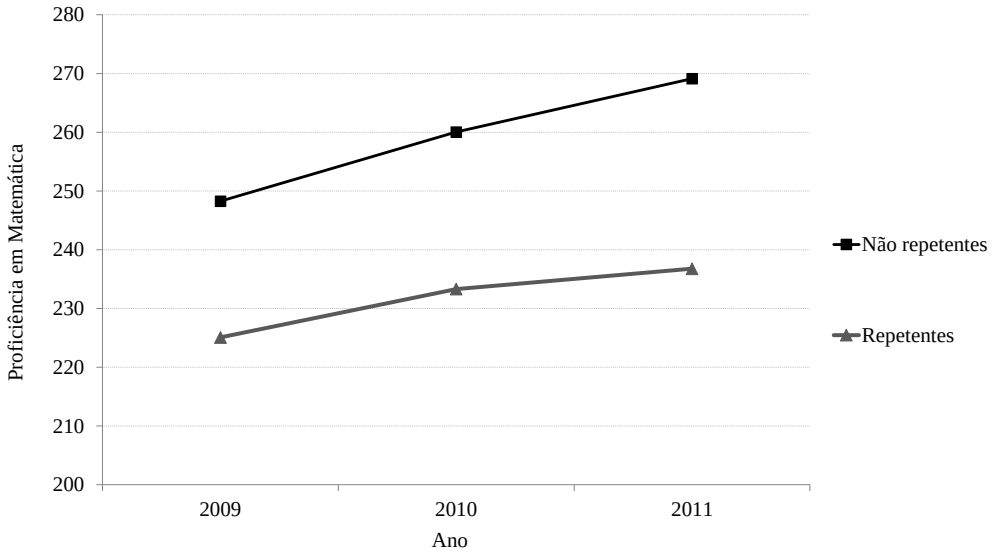

Figura 3
Proficiência média dos estudantes repetentes e não repetentes (n=7.549)

Em seguida foram realizadas as imputações dos dados de acordo com os procedimentos de imputação múltipla (IM) e mistura de padrões (MP). A Figura 4 apresenta as trajetórias estimadas utilizando esses procedimentos, as curvas em pontilhado correspondem às trajetórias estimadas. Pode-se observar que, para o padrão 2 e 3 de ausência, a trajetória estimada usando o procedimento IM apresenta estimativas mais elevadas para o desempenho se comparadas com as estimativas do procedimento MP. Esse resultado pode ser atribuído à abordagem adotada, uma vez que o procedimento MP supõe que os indivíduos com dados ausentes apresentam trajetória semelhante aos repetentes.

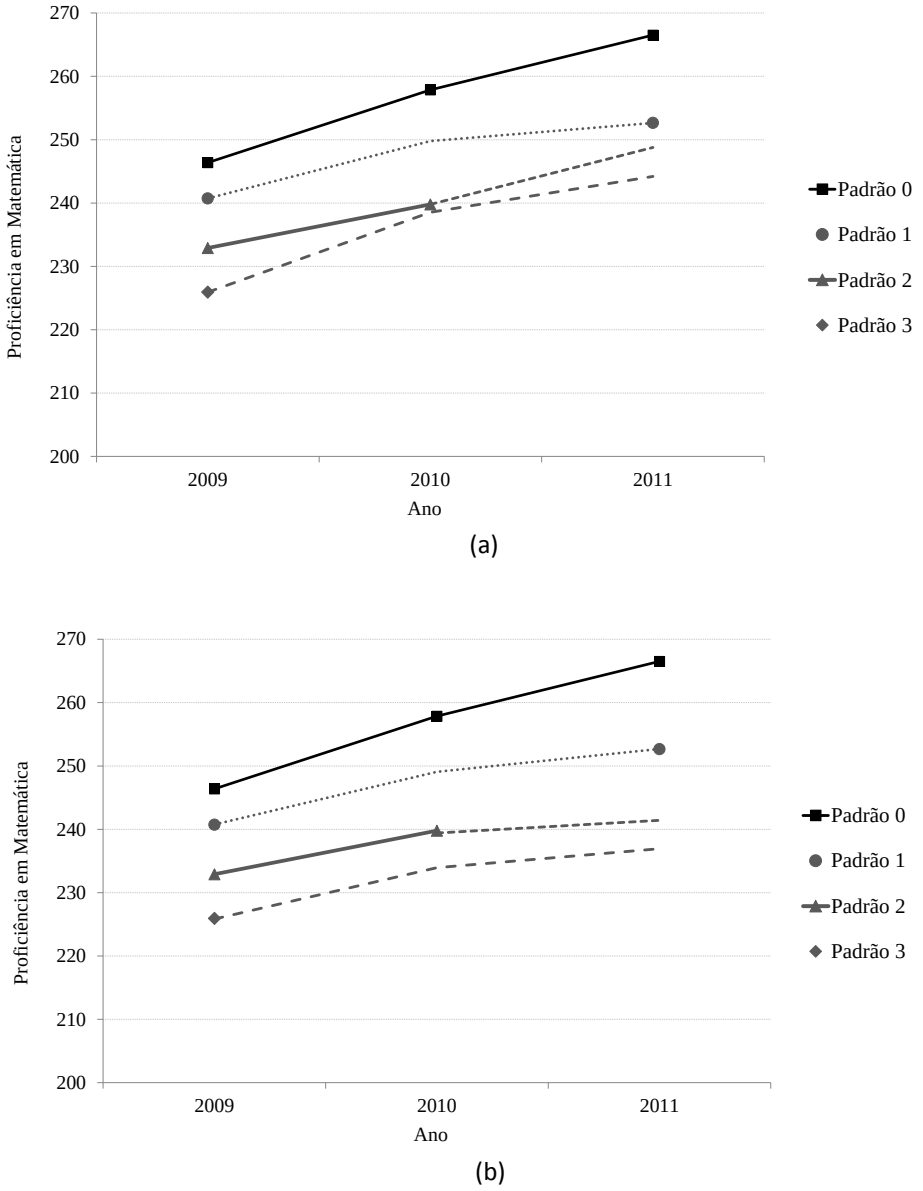

Figura 4
Proficiência média observada e estimada para os diferentes padrões de ausência – (a) procedimento IM e (b) procedimento MP

Por fim, na Tabela 4 são apresentados os resultados do ajuste do modelo de crescimento linear, utilizando os três procedimentos. Os procedimentos IM e MP apresentam intercepto menor, mas as diferenças observadas são relativamente pequenas, com variação inferior a 0,1% em relação ao valor 234,62 observado com o *listwise deletion* (LD). As estimativas das taxas médias de aprendizado também são menores, 9,57 pontos para o procedimento IM e 8,85 pontos para o procedimento MP, o que

corresponde a valores 9,5% e 16,4% menores em relação ao valor 10,58 obtido em LD, respectivamente.

Tabela 4
Comparação dos três procedimentos usando o modelo de crescimento linear

	LD			IM			MP		
Efeito fixo	Efeito	Erro Padrão	Razão	Efeito	Erro Padrão	Razão	Efeito	Erro Padrão	Razão
Intercepto	234,62	0,56	463,50	233,36	0,35	688,40	234,14	0,38	680,59
Ano	10,58	0,28	37,76	9,57	0,21	44,63	8,85	0,22	39,64
MAT08	0,66	0,01	67,67	0,65	0,01	86,82	0,64	0,01	86,09
Superior	7,47	0,85	8,75	9,29	0,66	14,02	9,24	0,66	13,92
Trabalho	-3,74	1,54	-2,45	-5,02	0,98	-5,13	-5,70	0,99	-5,76
Trabalho*ano	-0,79	0,91	-0,86	-1,84	0,60	-3,11	-1,80	0,61	-2,91
Efeito aleatório									
Intercepto	492,39			464,30			440,10		
Ano	23,44			29,09			33,37		
Erro	739,91			740,18			795,46		
N	5.162			8.681			8.681		

Pode-se observar que os procedimentos IM e MP apresentam efeitos mais acentuados para as variáveis *trabalho* e *superior*. O coeficiente estimado para a variável *superior* é aproximadamente 24% maior em relação ao resultado de LD, com estimativas iguais a 9,29 e 9,24, respectivamente. Para a variável *trabalho* observa-se que o procedimento IM apresenta estimativa igual a -5,02 para o efeito principal, um impacto 34,2% maior em relação ao obtido em LD (-3,47); no procedimento MP o efeito estimado para essa variável foi de -5,70, um aumento de 52,4% em relação à estimativa em LD. Nota-se também que a interação entre *trabalho* e *ano*, não significativa no procedimento LD, tem efeito significativo quando os procedimentos IM e MP são utilizados, com estimativas iguais a -1,84 e -1,80, respectivamente.

Considerando os efeitos aleatórios, verifica-se que a proficiência média no início do ensino médio apresenta maior variação no ajuste do procedimento LD. Por outro lado, o efeito do ano letivo apresenta maior variação entre os estudantes para os

procedimentos IM e MP. De forma geral, os erros padrão estimados são maiores no procedimento LD.

Discussão

Diversos métodos para o tratamento de dados ausentes são encontrados na literatura (Vinha, 2016a)[8]. A escolha do método adequado em cada situação é fundamental para mitigar os possíveis efeitos negativos da ausência de informação. Nesse sentido, a qualidade dos resultados obtidos a partir do uso de um determinado método depende diretamente da adequação das suposições relacionadas ao tipo de ausência. O objetivo do presente estudo foi apresentar uma comparação de três métodos baseados em suposições diferentes acerca do mecanismo gerador de ausência, tendo como base os dados longitudinais de uma avaliação educacional.

Os três procedimentos apresentaram estimativas distintas para os coeficientes do modelo proposto (Tabela 4). Essas diferenças podem ser explicadas pelas suposições impostas por cada abordagem. Por exemplo, assume-se que os dados ausentes são completamente ao acaso quando o procedimento LD é utilizado. Logo, por essa abordagem, desconsidera-se o fato de que os estudantes ausentes têm perfil diferente dos presentes e, em geral, o desempenho dos ausentes é inferior. Como consequência, esse procedimento apresenta a maior estimativa para a taxa média de aprendizado para os grupos com dados ausentes. Esses dados sugerem que o uso do procedimento LD leva a uma superestimação da taxa de aprendizado.

Com a utilização da IM assume-se que a ausência seria explicada pelo perfil dos estudantes e o desempenho observado nos períodos anteriores (ausência do tipo MAR). Com isso, o tratamento dos valores faltantes considera características do estudante ausente, assim os desempenhos estimados para esses indivíduos são mais próximos aos

[8] Manuscrito 1 apresentado neste documento.

observados para indivíduos com perfis semelhantes. Essa mudança no tratamento dos ausentes provavelmente é responsável pela redução da taxa média de aprendizado e o aumento do efeito das variáveis *trabalho* e *superior* em comparação ao procedimento LD, além do efeito significativo da interação entre *trabalho* e *ano*.

Por fim, o procedimento MP foi utilizado supondo que a ausência é do tipo MNAR. Nesse caso, a ausência não está apenas relacionada ao perfil dos estudantes e ao desempenho anterior, mas depende do valor a ser observado. Assumiu-se que a taxa de aprendizado dos estudantes ausentes é menor que a taxa observada para os presentes em geral. Como resultado, em relação à IM, verifica-se que a taxa média de aprendizado estimada é menor e o impacto da variável *trabalho* é maior.

A suposição de ausência completamente ao acaso é inadequada para os dados da avaliação do ensino médio cearense (Shirasu, 2014; Vinha, 2016b). Os indivíduos ausentes ou que abandonaram o ensino médio no Ceará têm perfis diferentes em termos de características socioeconômicas e do próprio desempenho escolar. Logo, os resultados obtidos a partir do procedimento LD devem ser analisados com cautela.

Como mencionado, não é possível avaliar empiricamente se os dados ausentes são do tipo MAR ou MNAR. Os dados perdidos não podem ser recuperados, logo não é possível avaliar a relação da ausência e esses valores. No entanto, é razoável supor que a ausência depende do desempenho a ser avaliado, principalmente nos casos em que a ausência é causada pelo abandono. Além disso, pela Figura 4, verifica-se que o tratamento dos dados pela IM estima trajetórias não esperadas para os grupos com ausentes. As trajetórias geradas pelo procedimento MP são mais coerentes neste contexto, os indivíduos com valores ausentes têm desempenho menor e taxa de aprendizado menor que os estudantes presentes em todo o estudo (Figura 4).

O presente estudo apresenta algumas limitações impostas pelos dados utilizados. Os bancos de dados disponibilizados contêm informações relativas a centenas de milhares de estudantes avaliados anualmente no ensino médio do Estado do Ceará, porém dada a dificuldade de interligação dos arquivos apenas uma pequena parcela dessa informação foi utilizada. Ainda, o procedimento de junção dos dados utilizado é passível de erro, dado que a avaliação é feita pela similaridade de nomes. Além disso, com os dados disponibilizados não é possível verificar se os indivíduos não avaliados em um ano estavam matriculados ou não, para obter essa informação seria necessária a utilização do Censo Escolar ou alguma outra fonte.

O procedimento MP proposto foi desenvolvido a partir dos resultados apresentados por outros autores (De Leon & Menezes-Filho, 2002; Neri, 2009; Shirasu, 2014; Silva, 2013; Simões, 2014; Soares & cols, 2015), que identificaram a relação entre o abandono e a repetência com o desempenho escolar e o perfil socioeconômico dos estudantes. A elaboração de um modelo baseado na mistura de padrões foi motivada pela busca de uma abordagem em que a suposição de identificação fosse clara e expressa em termos do contexto.

Vale ressaltar, no entanto, que outras abordagens para a análise de dados com valores ausentes do tipo MNAR poderiam ser aplicadas aos dados. Podem-se destacar os modelos de seleção e procedimentos baseados na reponderação da amostra (Fitzmaurice & cols, 2009). Além disso, outros modelos baseados na mistura de padrões poderiam ser utilizados, com a utilização de diferentes restrições de identificação.

Referências

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analysis. *Journal of School Psychology, 48*(1), 5-37.

Ceará (2011). Secretaria da Educação. SPAECE – 2011. *Boletim Pedagógico Matemática - Ensino Médio, 3, 1-22*.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 20*(10), 1-20.

Collins, L. M., Schafer, J. L., & Kam, C-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330-351.

De Leon, F. L. L., & Menezes-Filho, N. A. (2002). Reprovação, avanço e evasão escolar no Brasil. *Pesquisa e Planejamento Econômico, 32*(3), 417-452.

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable dropout. *Statistics in Medicine, 22*, 2553-2575.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods, 16*(1), 1-16.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2009). *Longitudinal data analysis.* Boca Raton: Chapman & Hall.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 115-142). New York: Springer-Verlag.

Gonçalves, M. E. (2008). *Análise de sobrevivência e modelos hierárquicos logísticos longitudinais: uma aplicação à análise da trajetória escolar (4ª a 8ª série - ensino*

fundamental). Tese de Doutorado, Universidade Federal de Minas Gerais, Belo Horizonte, MG.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206-213.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*, 64-78.

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychobiology, 45*(4), 1195-1199.

Karino, C. A., Vinha, L. G. A., & Laros, J. A. (2015). Os questionários do SAEB: O que eles realmente medem? *Estudos em Avaliação Educacional, 25*(59), 270-297.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88*, 125-134.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford Press.

Neri, M. (2009). *Motivos da evasão escolar*. Brasília: Fundação Getúlio Vargas.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525-556.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation, 7*(4), 353-383.

Rousseau, M., Simon, M., Bertrand, R., & Hachey, K. (2012). Reporting missing data: A study of selected articles published from 2003-2007. *Quality & Quantity, 46*(5), 1393-1406.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (second edition). Newbury Park: Sage.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Shirasu, M. R. (2014). *Determinantes da evasão e repetência escolar no Ceará*. Dissertação de Mestrado. Universidade Federal do Ceará, Fortaleza, CE, Brasil.

Silva, J. L. P. (2013). *Métodos de imputação múltipla para GEE em estudos longitudinais*. Dissertação de Mestrado, Universidade Federal de Minas Gerais, Belo Horizonte, MG.

Simões, A. (2014). *Acesso e evasão na Educação Básica: As perspectivas da população de baixa renda no Brasil*. Brasília: Ministério do Desenvolvimento Social e Combate à Fome.

Soares, T. M., Fernandes, N. S., Nóbrega, M. C., & Nicolella, A. C. (2015). Fatores associados ao abandono escolar no ensino médio público de Minas Gerais. *Educação e Pesquisa, 41*(3), 757-772.

Vinha, L. G. A. (2016a). *Dados ausentes em avaliações educacionais: Comparação de métodos de tratamento*. Manuscrito em preparação.

Vinha, L. G. A. (2016b). *Estudo de fatores associados ao desempenho escolar com dados ausentes não ignoráveis*. Manuscrito em preparação.

Apêndice 1 – Procedimento de junção dos arquivos de dados

Os dados utilizados no presente estudo foram disponibilizados pela Secretaria de Educação do Estado do Ceará em quatro conjuntos de arquivos, um para cada ano de avaliação (2008, 2009, 2010 e 2011). Cada conjunto era composto por um arquivo com as proficiências dos estudantes e outro com as respostas do questionário contextual. Os arquivos de dados de um mesmo ano têm um código comum de identificação dos estudantes. No entanto, a identificação dos estudantes muda de um ano para outro, logo a interligação de dois arquivos de anos diferentes não pôde ser feita diretamente (utilizando funções do tipo *merge* disponíveis nos pacotes estatísticos).

Essa situação é comum nos estudos em que duas ou mais fontes de informação são utilizadas. Nesses casos, diferentes bases de dados são interligadas por meio de técnicas de relacionamento de registros (*record linkage*). Essas técnicas utilizam campos como o nome, data de nascimento, nome dos pais, endereço e outras informações pessoais para relacionar os registros. Em geral, os procedimentos de *record linkage* encontrados na literatura são baseados no modelo proposto por Fellegi e Sunter (1969). Por esse modelo, são gerados escores de concordância para cada par de registros e, de acordo com os valores obtidos para o par, decide-se se os registros são ou não do mesmo indivíduo (Cabral, 2010).

No presente estudo foi utilizado um procedimento semelhante ao proposto por Fellegi e Sunter (1969). No entanto, como os arquivos não continham outras informações pessoais além de nome e sexo do estudante, não seria eficiente estimar escores de concordância tendo como base apenas dois campos. Assim, uma medida de similaridade de cadeias de caracteres aplicada ao campo do nome foi utilizada como critério para identificação de possíveis pares de registros.

O procedimento foi aplicado em uma amostra composta por 11.245 estudantes do primeiro ano do ensino médio selecionada aleatoriamente do arquivo de proficiências de 2009 (o que corresponde a cerca de 10% dos estudantes registrados nesse ano). Entre esses alunos, 226 não responderam ao questionário contextual, logo a amostra utilizada

como base é composta por 11.019 alunos com as proficiências em matemática e língua portuguesa e com as respostas do questionário.

A interligação dos arquivos seria feita diretamente através de funções do tipo *merge* presentes nos pacotes estatísticos ou planilhas, caso os nomes dos estudantes fossem registrados sem qualquer erro. No entanto, esse registro pode apresentar diversas variações para o mesmo nome, como consequência dos erros de leitura (os nomes foram escritos pelos alunos e decodificados por leitura ótica), uso de abreviações e omissão de parte do nome. Logo, fez-se necessária a utilização de critérios de similaridade de nomes para a identificação dos estudantes presentes em 2009 nos arquivos dos outros anos.

A medida de similaridade de nomes utilizada foi gerada pela função *compged* do SAS. Essa função gera uma medida de diferença entre dois conjuntos de caracteres, essa medida depende do número de mudanças necessárias para que um conjunto de caracteres fique igual ao outro. A função possibilita a atribuição de diferentes pesos para as operações realizadas (substituição, retirada, inserção de letra, por exemplo), porém optou-se pela utilização de pesos iguais para todas as mudanças realizadas. Quando os nomes são exatamente iguais a função *compged* assume o valor 0, qualquer diferença observada entre os nomes é penalizada, e o valor observado é maior que 0. Por exemplo, na comparação dos nomes “LUIS GUSTAVO VINHA” e “LUIS GVSTAVO VINHAS” a função assume o valor 2, pois foram necessárias duas mudanças no primeiro nome para igualá-lo ao segundo (uma mudança e uma retirada de letra).

A seguir são apresentados os passos realizados na junção dos registros dos estudantes em 2009 e 2010.

Passo 1: Identificação dos possíveis pares usando a função *compged*

Para cada registro da amostra de 2009 foram identificados os quatro possíveis pares em 2010. Esses possíveis pares são os registros do arquivo de proficiência de 2010 que possuíam maior semelhança de nomes (usando a função *compged*). Como resultado foi gerado um arquivo com os nomes, código da escola e algumas outras informações dos

estudantes em 2009 e as informações relativas aos quatro registros com nomes mais próximos de 2010.

Passo 2: Análise dos possíveis pares e divisão em grupos

O arquivo gerado no primeiro passo foi então analisado. Os registros foram então divididos em: *pares na mesma escola*; *pares com mudança de escola, ausente em 2010* e *dúvida*. Vale ressaltar que a divisão dos registros foi feita caso a caso.

1) *Pares na mesma escola*. Esse conjunto é formado por pares de registros, um registro de 2009 e um de 2010, com nomes iguais ou muito semelhantes, na mesma escola. Nos casos em que os nomes não eram iguais, mas muito próximos e com o mesmo código da escola, muitos registros foram considerados pares uma vez que a diferença observada parecia ser relativa a um erro de leitura. A seguir são apresentados alguns exemplos[9]:

   - Apenas uma letra trocada:

     "LVIS GUSTAVO AMARAL DO VINHA" e

     "LUIS GUSTAVO AMARAL DO VINHA";

   - Ausência de espaço entre os nomes:

     "LUIS GUSTAVO DO AMARAL VINHA" e

     "LUIS GUSTAVO DO AMARALVINHA";

   - Uma letra faltando:

     "LUIS GUSTAVO DO AMARAL VINHA" e

     "LUIS GUSTAVO DO AMARAL VIHA";

   - Sobrenomes abreviados:

     "LUIS GUSTAVO DO AMARAL VINHA" e

     "LUIS GUSTAVO A VINHA";

[9] Nos exemplos não são usados os nomes dos alunos para que seja mantido o anonimato. As diferenças apresentadas são iguais às observadas no banco de dados real.

- Primeiro nome abreviado:

  "L GUSTAVO DO AMARAL VINHA" e

  "LUIS GUSTAVO DO AMARAL VINHA";

2) *Pares com mudança de escola*. Neste passo foi formado um conjunto de pares de registros que não pertenciam ao conjunto anterior e que tinham nomes iguais ou muito semelhantes, mas estavam escolas diferentes. O procedimento adotado neste passo foi o mesmo do anterior, mas nesse caso foram considerados os possíveis pares de outras escolas;
3) *Ausentes em 2010*. Todos os registros de 2009 que não têm registro com nome igual ou semelhante em 2010 foram incluídos nesse conjunto. Em geral, os registros têm valor da função *compged* elevado para os possíveis pares de 2010, em geral acima de 5;
4) *Dúvida*. Para uma parcela dos registros de 2009 não foi possível classificar como pertencentes a um dos três grupos descritos acima. Entre esses casos estão:
   - Registro de 2009 com nome igual ou semelhante a dois ou mais registros de 2010;
   - Registro de 2009 com nome muito semelhante a um registro de 2010, mas a diferença sugere que sejam indivíduos diferentes. Por exemplo:

     "DIEGO AMARAL VINHA" e

     "DIOGO AMARAL VINHA" ou

     "MARIA AMARAL VINHA" e

     "MARIO AMARAL VINHA";

Passo 3: Seleção dos pares a serem utilizados nas análises

Ao final do passo anterior foi criado um novo arquivo contendo o resultado da identificação e divisão dos registros. Esse novo arquivo contém o grupo e a identificação dos estudantes em 2009 e em 2010. Neste estudo foram utilizados apenas os registros classificado como *pares na mesma escola* e *ausente em 2010*.

O mesmo procedimento foi utilizado para o pareamento dos arquivos de 2011 e 2008. Para o arquivo de 2008 o código da escola perde a importância uma vez que a mudança da escola é comum na passagem do ensino fundamental para o médio, nesse caso apenas o nome dos estudantes foi considerado e a classificação pares com mudança de escola não foi usada.

Por fim, um arquivo contendo o resultado dos procedimentos foi consolidado. Neste arquivo final todos os estudantes presentes na amostra de 2009 foram classificados como ausente ou presente em 2008, 2010 e 2011.

Referências

Cabral, M. D. B. (2010). *Proposta de relacionamento probabilístico dos registros da base de dados do programa de rastreamento do câncer do colo do útero*. Tese de Doutorado, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*, 1183-1210.

Manuscrito 3

ESTUDO DE FATORES ASSOCIADOS AO DESEMPENHO ESCOLAR COM DADOS AUSENTES NÃO IGNORÁVEIS

Resumo

O presente estudo teve como objetivo identificar os fatores associados ao desempenho em matemática de estudantes do ensino médio cearense. Para isso, foram utilizados os dados do acompanhamento realizado pelo SPAECE no período de 2009 a 2011. Dos 8.681 estudantes selecionados 25,7% estava ausente em pelo menos um momento da avaliação. Constatou-se que essa ausência estava relacionada às características dos estudantes e ao desempenho escolar avaliado. Por esse motivo, foi utilizado um método desenvolvido para o tratamento de dados ausentes não ao acaso (MNAR). Os resultados do ajuste do modelo de crescimento linear mostraram que a taxa anual de aprendizado na prova de matemática foi de 8,96 pontos em média, e que essa taxa varia significativamente entre os estudantes. Com a utilização de dados longitudinais e técnicas de tratamento de dados ausentes, os resultados corroboram estudos transversais de fatores associados ao desempenho escolar. Além disso, demonstra que variáveis relacionadas à idade, número de reprovações e período noturno têm efeitos negativos tanto na proficiência inicial, quanto na taxa de aprendizado.

Palavras-chave: SPAECE, desempenho escolar, tratamento de dados ausentes, dados ausentes não ao acaso, modelo de crescimento linear.

Abstract

This study aimed to identify factors associated with mathematics performance of high school students in the State of Ceará, Brazil. The SPAECE database from 2009 to 2011 was used. Of the 8,681 selected students 25.7% were absent at least one moment of follow-up. The missingness was found to be related to students' characteristics and school performance. Thus, a method developed for the treatment of missing not random (MNAR) data was used. Results of the adjustment of the linear growth model showed that the annual learning rate on the mathematics test was 8.96 points on average, and this rate varies significantly among students. Using longitudinal data and missing data treatment techniques, the results corroborate those from cross-sectional studies of factors associated with school performance. Moreover, it shows that variables related to age, school repetition and evening classes have negative effects on both initial proficiency and learning rate.

Keywords: educational assessment, school performance, missing data treatment, missing not at random, linear growth models.

Introdução

A educação no Brasil avançou muito nas últimas décadas, principalmente no que se refere ao acesso à escola. Entretanto, o país ainda apresenta indicadores educacionais bastante desfavoráveis, tanto em relação à qualidade do ensino oferecido, quanto ao fluxo dos estudantes no sistema (Franco & cols, 2007).

Entre os diversos problemas a serem enfrentados na busca pela melhoria da educação no país, os maiores desafios estão relacionados ao ensino médio[10]. Com base nos dados do SAEB (Sistema de Avaliação da Educação Básica) pode-se verificar que os estudantes do ensino médio têm os piores desempenhos entre as três etapas avaliadas[11]. Segundo o movimento "Todos pela Educação", em 2013, apenas 27,2% dos estudantes no 3º ano do ensino médio tinham o nível de proficiência esperado em língua portuguesa. Quando considerada a proficiência em matemática, o quadro era ainda pior, somente 9,3% dos jovens brasileiros terminaram essa etapa com o conhecimento adequado na disciplina. Além disso, analisando o histórico desses percentuais, observou-se uma estagnação entre os anos de 1999 e 2009, seguida de uma leve tendência de queda nos últimos anos (Todos pela educação, 2015).

O ensino médio brasileiro enfrenta também o desafio de manter os jovens nas escolas. Em 2011, de acordo com os dados da PNAD (Pesquisa Nacional por Amostra de Domicílios), mais de 1,7 milhão de adolescentes de 15 a 17 anos não estavam estudando, o que corresponde a 16,3% dessa população (Volpi, Silva, & Ribeiro, 2014). Além disso, verificou-se que o percentual de jovens fora da escola era maior para a população de baixa renda e entre estudantes autodeclarados pretos e pardos.

[10] "O ensino médio é o maior desafio da educação do País". Maria de Salete Silva, coordenadora do programa de educação do Fundo das Nações Unidas para a Infância, no Brasil (Unicef) em setembro de 2013. Disponível em: *http://www.istoe.com.br/reportagens/326686_O+MAIOR+PROBLEMA+DA+EDUCACAO+DO+BRASIL*

[11] O SAEB avalia também os estudantes do 5º e 9º ano do ensino fundamental.

Trata-se de um quadro alarmante, dada a importância da educação na vida dos jovens e no desenvolvimento do país. O retorno da educação no Brasil é conhecido e largamente estudado (Neri, 2009). Sabe-se que quanto maior o grau de escolaridade de um indivíduo maior é sua empregabilidade e renda, o que reflete positivamente em outras dimensões, como bem-estar e saúde. Ainda, o mercado de trabalho atual exige maior formação dos trabalhadores, uma vez que os processos produtivos foram profundamente transformados pelas inovações tecnológicas (Krawczyk, 2011).

O contexto atual mostra a necessidade de ações para a melhoria da educação oferecida aos adolescentes brasileiros. Nesse sentido, para que essas ações sejam efetivas, é necessário que as políticas educacionais estejam embasadas em resultados de pesquisas (Vinha, Karino, & Laros, no prelo). Essas pesquisas podem, a partir dos dados gerados pelas avaliações educacionais, identificar fatores associados à trajetória escolar e assim subsidiar ações e programas de melhoria.

As avaliações educacionais em larga escala realizadas no Brasil têm sido usadas nos estudos de eficácia escolar, principalmente como ferramenta para identificação de fatores associados ao desempenho acadêmico, mas também para análise do efeito escola e avaliações de programas (Karino & Laros, 2016; Vinha & Laros, 2016). Em geral, são utilizados os dados dos estudantes avaliados no SAEB tendo como indicadores de desempenho as proficiências em matemática e língua portuguesa, além das informações obtidas pelos questionários contextuais.

O presente manuscrito visa a contribuir com a identificação dos fatores associados ao desempenho de estudantes do ensino médio. Este estudo inova com a análise de dados de uma avaliação longitudinal e com a utilização de uma nova abordagem que considera a relação existente entre desempenho, abandono e evasão escolar. A escolha

da abordagem e das variáveis utilizadas nas análises baseia-se nos resultados de pesquisas anteriores, a seguir são apresentados alguns desses resultados.

Estudos de fatores associados ao desempenho escolar no ensino médio

O estudo de Soares, Alves e Oliveira (2001) utilizou os dados do vestibular da Universidade Federal de Minas Gerais (UFMG) para avaliar o efeito das escolas e fatores associados ao desempenho dos estudantes do ensino médio mineiro. O efeito escola estimado foi de 33% e com a introdução das variáveis dos alunos no modelo o efeito se aproxima dos 20%. Verificou-se que sexo, nível socioeconômico, turno, habilidade anterior e a forma de preparo para o vestibular são importantes preditoras do desempenho no exame. No entanto, os autores ressaltam que a amostra utilizada tem viés de seleção devido ao critério utilizado para inclusão das escolas, e que isso deve ser considerado na interpretação dos resultados.

Andrade e Laros (2007) estudaram os fatores associados ao desempenho em matemática e língua portuguesa utilizando os dados do SAEB de 2001. A modelagem utilizada considerou dois níveis de agregação, nível do aluno e da escola. O efeito escola estimado foi igual a 46%, e após a inclusão das variáveis de controle passou para 17%. Os autores destacaram a forte relação entre desempenho e atraso escolar, além do efeito significativo das variáveis de controle (nível socioeconômico, sexo e etnia), recursos culturais, relação da família com a escola, dever de casa e trabalho, e também de algumas características das escolas.

Gonçalves e França (2008) mostraram que as desigualdades sociais são reproduzidas pelo sistema de educação básica no Brasil. Eles utilizaram os dados da edição de 2003 do SAEB e outras informações disponibilizadas pelo Instituto de Pesquisa Econômica Aplicada (IPEA). Por meio de modelos hierárquicos com três níveis de agregação (estudantes, escolas e estados) os autores verificaram que as escolas

eram responsáveis por 42,3% da variabilidade das notas dos estudantes em matemática no ensino médio. Entre outros resultados os autores apontam a influência do sexo, etnia, nível socioeconômico e número de reprovações no desempenho escolar.

O objetivo do estudo realizado por Laros, Marciano e Andrade (2012) era identificar as características de alunos e escolas associadas ao desempenho em língua portuguesa. Além disso, os autores visavam verificar as diferenças de desempenho entre as regiões do país, com base nos dados SAEB 2001. O efeito escola estimado estava em torno de 33% sem o uso das variáveis de controle, e 12% quando essas variáveis foram introduzidas no modelo. Nesse estudo, destacou-se o impacto do atraso escolar e dos recursos culturais no desempenho em língua portuguesa. Na comparação entre as regiões do país, algumas diferenças foram observadas, entre elas, verificou-se que o impacto do nível socioeconômico médio da escola é maior na Região Nordeste e menor na Região Sul.

Além dos problemas relacionados ao aproveitamento dos estudantes, o ensino médio também enfrenta o desafio de manter os jovens nas escolas. Estudos relacionados à evasão e ao abandono escolar também são encontrados na literatura, a seguir são apresentados alguns desses estudos.

Estudos dos fatores associados à evasão e ao abandono escolar

De Leon e Menezes-Filho (2002) identificaram os fatores relacionados a reprovação, avanço e evasão escolar no período de 1984 a 1997. A partir dos dados da PME (Pesquisa Mensal de Emprego) os autores identificaram que, para os estudantes com reprovação, a probabilidade de evasão escolar é maior entre os mais velhos e os que moram sem os pais, por outro lado essa probabilidade é menor com o aumento da escolaridade do chefe da família.

Com base nos dados da PNAD de 2006, Neri (2009) verificou que a falta de interesse dos estudantes e dos pais pela educação é a principal causa da evasão escolar dos jovens de 15 a 17 anos, sendo identificada por 40,3% das famílias entrevistadas. A falta de renda e a dificuldade de acesso à escola foram outros fatores apontados pelo autor, com 27,3% e 10,9% dos casos, respectivamente.

Shirasu (2014) estudou os fatores associados à evasão e à repetência escolar dos estudantes do ensino médio de escolas públicas do estado do Ceará. Para isso foram utilizados os dados do SPAECE (Sistema Permanente de Avaliação da Educação Básica do Ceará) no período de 2008 a 2011. A autora afirma que a taxa de evasão é maior entre os estudantes do sexo masculino, estudantes com atraso ou que foram reprovados em anos anteriores. Ainda, o estudo indica que a escolaridade dos pais, participação no programa Bolsa Família e hábito de fazer as tarefas de casa estão negativamente associados à taxa de evasão.

Segundo o relatório "10 desafios do ensino médio" lançado pelo UNICEF[12] em 2014, os jovens brasileiros enfrentam inúmeros obstáculos para permanecer no ensino médio, entre eles: condição socioeconômica, trabalho precoce, gravidez, violência familiar e no entorno. Além desses há fatores relacionados às escolas, como organização, conteúdos distantes da realidade, falta de diálogo entre os atores (alunos, professores e gestores), desmotivação de professores e problemas de infraestrutura. Esse estudo qualitativo contou com a participação de 250 adolescentes de seis cidades brasileiras, com a realização de grupos focais e entrevistas em profundidade (Volpi, Silva, & Ribeiro, 2014).

A pesquisa realizada por Soares, Fernandes, Nóbrega e Nicolleta (2015) utilizou os dados da PNAD e da Pesquisa sobre Abandono Escolar (PSAE) realizada no estado

[12] Fundo das Nações Unidas para a Infância.

de Minas Gerais. Os autores confirmaram que o abandono escolar é influenciado por fatores já conhecidos e apontados na literatura como sexo, defasagem idade/série, condição socioeconômica e a necessidade de trabalhar para ajudar a família. Eles observaram também a influência da gravidez durante o ensino médio, composição familiar e variáveis relacionadas ao contexto da escola. Vale ressaltar a relação entre o desempenho acadêmico e o abandono identificada no estudo, pois os estudantes que apresentavam maior dificuldade nas disciplinas tinham maiores chances de abandonar a escola.

Análise dos dados de avaliações educacionais

As avaliações em larga escala têm papel fundamental no processo de melhoria da educação. Contudo, a análise dos dados deve considerar algumas características desse tipo de levantamento. A estrutura da população de interesse, o período de coleta dos dados (levantamentos transversais ou longitudinais) e a ocorrência de dados ausentes devem ser considerados na escolha das técnicas utilizadas e na interpretação dos resultados.

A análise dos dados de avaliações educacionais deve considerar a estrutura hierárquica da população de interesse (Lee, 2008; Vinha, Karino, & Laros, no prelo). Os alunos de uma turma ou escola compartilham experiências, têm os mesmos professores, utilizam os mesmos materiais e equipamentos e, em geral, residem no mesmo bairro. Por sua vez, as escolas que pertencem a uma rede de ensino têm métodos, perfis de professores e estruturas semelhantes e assim por diante. A utilização dos modelos multinível (ou modelos de regressão hierárquica) é recomendada nesses casos (Goldstein, 2010; Hox, 2010; Raudenbush & Bryk, 2002).

Muitos estudos publicados na literatura brasileira utilizam os dados transversais do SAEB com a aplicação de modelos multinível (Karino & Laros, 2016). Contudo, o uso

desses dados não possibilita a identificação de relações de causa e efeito. Se a análise é realizada a partir da proficiência e condições escolares específicas em um determinado ano, os resultados podem ser frágeis e não reveladores, uma vez que o aprendizado é um processo acumulativo, que reflete a trajetória escolar do aluno (Franco, 2001).

A rigor, o efeito da escola no aprendizado dos alunos só poderia ser estimado com dados longitudinais (Soares & Candian, 2007). Nesse tipo de levantamento os mesmos alunos são avaliados em diferentes períodos de tempo e, dessa forma, é possível identificar os fatores que influenciam a evolução desses estudantes. Entretanto, avaliações educacionais em larga escala com enfoque longitudinal ainda são raros no Brasil. Pode-se destacar o estudo Geres (Estudo Longitudinal da Geração Escolar 2005) onde uma amostra de estudantes do ensino fundamental de escolas públicas e particulares de cinco cidades brasileiras foi avaliada em quatro anos consecutivos. Outro exemplo é o acompanhamento dos estudantes do ensino médio no estado do Ceará realizado pelo SPAECE.

Outro fator importante que deve ser considerado é a perda de informação. A ocorrência de valores ausentes é constante na pesquisa educacional em larga escala (Cheema, 2014), em especial nos levantamentos longitudinais.

A utilização de dados incompletos gera necessariamente uma redução na precisão dos resultados da pesquisa, qualquer que seja o padrão da informação perdida. Entretanto, mais grave que a perda de precisão é a possível introdução de viés nas estimativas, o que por sua vez pode gerar conclusões enganosas acerca do fenômeno estudado. A escolha da técnica de análise apropriada é fundamental para minimizar os efeitos da ausência de informação (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009; Schafer & Graham, 2002).

A escolha do método para tratamento dos dados está diretamente relacionada ao tipo de ausência observado no estudo. De acordo com a classificação proposta por Rubin (Rubin, 1976; Rubin, 1987), os dados ausentes podem ser gerados por três mecanismos: completamente ao acaso (MCAR - *Missing Completely at Random*); ao acaso (MAR - *Missing at Random*); e não ao acaso, (MNAR - *Missing not at Random*).

Os dados são classificados como MCAR quando a taxa de ausência não está relacionada a qualquer variável presente no estudo. Esse tipo de ausência tem o menor impacto nas análises e nos resultados, uma vez que a amostra de valores completos (desconsiderando as observações com dados faltantes) pode ser vista como uma amostra representativa da população (Vinha, 2016). Com isso, as técnicas estatísticas tradicionais podem ser usadas sem a introdução de viés nas estimativas.

A ocorrência de dados ausentes do tipo MAR está associada a variáveis observadas no estudo. Nesse caso, a amostra de valores completos não representa a população de interesse, o que pode gerar viés na estimação de parâmetros populacionais. No entanto, a análise de dados com ausência do tipo MAR pode gerar inferências válidas quando são utilizadas técnicas adequadas. Esse tipo de ausência também é considerado ignorável, uma vez que não se faz necessária a modelagem adicional do mecanismo de ausência de dados no processo de estimação (Fitzmaurice & colaboradores 2009).

Os dados MNAR ocorrem quando a ausência está associada aos valores observados no estudo e também aos valores não observados. Esses dados também são chamados de não ignoráveis, pois o mecanismo gerador de ausência deve ser modelado para que sejam obtidas boas estimativas dos parâmetros de interesse (Allison, 2001).

Procedimentos tradicionais de tratamento de dados ausentes, como o *listwise deletion*, ainda são muito usados na pesquisa social aplicada (Cheema, 2014; Peugh &

Enders, 2004; Rousseau, Simon, Bertrand, & Hachey, 2012). No entanto, esses métodos são criticados, pois podem apresentar bons resultados somente quando os dados ausentes são completamente ao acaso (MCAR), ou em algumas situações específicas (Enders, 2010).

O desenvolvimento de métodos mais sofisticados para análise de dados com observações ausentes tem sido intenso nas últimas décadas. Os métodos baseados na estimação por máxima verossimilhança e imputação múltipla têm recebido atenção especial por parte dos pesquisadores. Em geral, esses métodos apresentam melhores desempenhos que as técnicas tradicionais, principalmente quando os dados ausentes são do tipo MAR (Graham, 2009). A literatura relacionada aos métodos de tratamento de dados MNAR também é grande, principalmente na área de bioestatística (Enders, 2011). Destacam-se os modelos de seleção e de mistura de padrões (Fitzmaurice & colaboradores 2009), e os procedimentos baseados na reponderação da amostra (Schafer e Graham, 2002).

O SPAECE e a evolução da educação no Estado do Ceará

O Estado do Ceará tem papel importante no cenário nacional no que se refere à avaliação da educação. Em 1992 foi lançada a primeira edição do SPAECE, sistema que avalia censitariamente os alunos matriculados no ensino fundamental e médio da rede pública. O SPAECE tem como objetivo fornecer subsídios para formulação, reformulação e monitoramento das políticas educacionais, e ainda visa apresentar a professores, dirigentes escolares e gestores um quadro da situação da Educação Básica da rede pública de ensino do estado (http://www.spaece.caedufjf.net/o-programa/).

O sistema tem três focos de atuação. Primeiro, o nível de proficiência em leitura dos estudantes é avaliado no 2º ano do ensino fundamental. Segundo, ainda no ensino fundamental, o sistema avalia as competências e habilidades dos alunos nas disciplinas

de língua portuguesa e matemática no final de cada etapa (5º e 9º ano). O terceiro foco está no monitoramento dos três anos do ensino médio, nesse caso avalia-se o progresso de aprendizagem dos alunos por se tratar de uma avaliação longitudinal.

O estado do Ceará tem investido em programas de melhoria da educação e os resultados podem ser observados em alguns indicadores (Ribeiro, Júnior, & Lourenço, 2013). A partir da análise dos dados do SAEB de 2005 e 2011, é evidente o avanço da qualidade do ensino oferecido no Estado do Ceará, com destaque para os resultados observados no ensino fundamental. Nesse período, considerando os resultados dos testes de proficiência em matemática e língua portuguesa, observou-se que o Ceará apresenta a maior evolução quando comparado com os outros estados e estava entre os treze estados com melhor desempenho em 2011(IPECE, 2012). Além do avanço em termos de desempenho, observou-se no mesmo período o aumento das taxas de aprovação entre os estudantes cearenses (Ribeiro, Júnior, & Lourenço, 2013). No entanto, o ensino médio não apresenta a mesma evolução observada nas demais etapas da educação básica, entre outros resultados vale destacar uma leve queda no desempenho médio dos estudantes em matemática entre 2005 e 2011.

O presente estudo utiliza os dados do acompanhamento dos estudantes do ensino médio realizado pelo SPAECE, com o objetivo de identificar os fatores associados ao desempenho acadêmico desses estudantes. Mais especificamente, o estudo visa identificar as características dos estudantes e suas famílias que estão relacionadas com a proficiência em matemática no início do ciclo, e com a evolução desses indivíduos ao longo do período avaliado. O método escolhido para as análises considera a natureza longitudinal da avaliação e a presença de dados ausentes não ignoráveis. O estudo também tem como objetivo ressaltar a necessidade da análise do tipo de dados ausentes e da utilização de métodos que considerem o tipo de ausência observado.

Método

Participantes

Este estudo utiliza os dados do SPAECE relativos ao acompanhamento do ensino médio realizado no período de 2009 a 2011. A amostra foi selecionada entre os estudantes matriculados no primeiro ano do ciclo em 2009, avaliados por meio do teste de Matemática e que responderam ao questionário socioeconômico. A interligação dos dados das proficiências avaliadas em 2010 e 2011 foi feita usando o nome do aluno como identificador, uma vez que os arquivos não contêm uma identificação numérica comum para os estudantes nos diferentes anos. A descrição do procedimento utilizado para interligação dos bancos de dados é apresentada em Vinha (2016)[13]. Os 8.681 estudantes selecionados estavam matriculados em 534 escolas, distribuídas em 184 municípios.

Instrumentos

Nas avaliações do SPAECE, o desempenho escolar dos estudantes é medido por meio de testes cognitivos compostos por itens elaborados com base nos parâmetros curriculares do Ministério da Educação. As proficiências são calculadas pela teoria de resposta ao item, o que permite a comparação do desempenho dos estudantes em diferentes períodos. A escala utilizada é a mesma escala de proficiência do SAEB, com média 250 e desvio padrão 50, tendo como referência o 9º ano do Ensino Fundamental de 1997 (Ceará, 2011). Questionários contextuais também são utilizados, esses instrumentos são compostos por itens relacionados ao nível socioeconômico e cultural das famílias, práticas de estudos e clima em sala de aula. As variáveis utilizadas no estudo são apresentadas na Tabela 1.

[13] Manuscrito 2 apresentado neste documento.

Tabela 1
Variáveis utilizadas no estudo.

Variáveis	**Descrição / Codificação**
MAT09	Desempenho na prova de matemática em 2009.
MAT10	Desempenho na prova de matemática em 2010.
MAT11	Desempenho na prova de matemática em 2011.
MAT08	Desempenho na prova de Matemática em 2008.
Sexo	Masculino (1) e feminino (0).
Etnia	Variável indicadora que assume o valor 1 para o estudante autodeclarado branco ou amarelo e 0 nos demais casos.
Turno	Turno em que o estudante frequenta as aulas: manhã, tarde ou noite.
Tarde	Variável indicadora que assume o valor 1 se o estudante frequenta a escola no período vespertino.
Noite	Variável indicadora que assume o valor 1 se o estudante frequenta a escola no período noturno.
Idade	Idade reportada em 2009 (em anos).
Idade_i	Variável indicadora que assume o valor 1 se o estudante declara ter 17 anos ou mais no primeiro ano do ensino médio e 0 se declara ter menos de 17 anos.
Escolaridade da mãe	Nunca estudou; 1ª a 4ª série do ensino fundamental; 5ª e 8ª série do ensino fundamental; 1ª e a 3ª série do ensino médio; cursou ensino superior; não sabe.
Escolaridade_i	Variável indicadora que assume 1 se a mãe estudou pelo menos até a primeira série do ensino fundamental e 0 se não estudou ou se o indivíduo não sabe o grau de escolaridade da mãe.
Reprovações	Número de reprovações reportado em 2009: nunca repetiu; uma reprovação; duas reprovações; três ou mais reprovações.
Reprovações_i	Variável indicadora que assume o valor 1 se o estudante afirma ter sido reprovado pelo menos uma vez e 0 quando o estudante não havia sido reprovado até 2009.
Itens	Índice obtido pela análise de componentes principais a partir dos seguintes itens presentes na casa da família do estudante: banheiro, rádio, geladeira, televisão em cores, máquina de lavar roupa, aparelho de DVD, automóvel.
Superior	Variável indicadora que assume o valor 1 quando o estudante pretende ingressar no ensino superior e 0 se tem outros planos após a conclusão do ensino médio.
Trabalho	Variável indicadora que assume o valor 1 se o estudante trabalha fora de casa e 0 caso contrário.

Análise de dados

As análises foram realizadas utilizando o software estatístico SAS (*Statistical Analysis System*), versão 9.4. O tratamento de dados ausentes utilizado neste estudo foi

proposto por Vinha (2016). Esse procedimento baseia-se nos modelos de misturas de padrões (Enders, 2011) e na imputação múltipla (Allison, 2001), sob a suposição de que a ausência de informação não é ao acaso (MNAR). Nesse caso, foram gerados 100 bancos de dados com valores imputados segundo o procedimento proposto e, para cada banco de dados, foi ajustado o modelo de interesse. As estimativas finais são combinadas utilizando o método proposto por Rubin (1987).

O estudo baseia-se na estimação do modelo de crescimento linear (Raudenbush & Bryk, 2002). Esse modelo é apresentado considerando uma estrutura em dois níveis: o primeiro nível corresponde às ocasiões e o segundo é o nível dos indivíduos. Dado o número reduzido de observações para cada estudante, o componente temporal utilizado é descrito por uma função do primeiro grau. A equação que descreve o primeiro nível é dada por

$$Y_{ij} = \pi_{0j} + \pi_{1j}(ANO)_i + \varepsilon_{ij},$$

onde Y_{ij} é a proficiência do aluno j, no ano i; π_{0j} é a proficiência esperada do aluno j no início do ensino médio; π_{1j} é a taxa de aprendizado do aluno j em um ano acadêmico; e ε_{ij} é a parcela aleatória da proficiência não explicada pelo modelo.

Os coeficientes π_{0j} e π_{1j} podem variar de acordo com as características dos alunos e suas famílias. Por exemplo, uma medida de proficiência anterior e a idade dos estudantes podem ser usadas como variáveis independentes no nível dos alunos. As equações para esse nível são:

$$\pi_{0j} = \beta_{00} + \beta_{01}(Profiência\ anterior)_j + \beta_{02}(Idade)_j + r_{0j} \text{ e}$$

$$\pi_{1j} = \beta_{10} + \beta_{11}(Idade)_j + r_{1j}.$$

Por essa formulação, a proficiência anterior tem influência na proficiência inicial e a variável idade é importante tanto para a proficiência inicial quanto para a taxa de

aprendizado. Nesse caso, o coeficiente β_{01} representa a variação no desempenho inicial com o aumento de uma unidade na proficiência anterior. O coeficiente β_{11} quantifica a mudança na taxa anual de aprendizagem com a variação da idade. Os termos r_{0j} e r_{1j} correspondem aos componentes aleatórios da proficiência esperada e da taxa de aprendizado.

A proficiência em matemática foi utilizada como variável dependente no modelo de crescimento linear. A construção do modelo segue os passos propostos por Hox (2010):

1. Ajuste do modelo nulo. Esse modelo contém apenas o intercepto e as variâncias nos níveis do aluno e da ocasião. Os resultados são utilizados no cálculo da correlação intraclasse.
2. Inserção da variável relativa ao tempo. Essa variável assume os valores 1, 2 e 3, para as avaliações de 2009, 2010 e 2011, respectivamente. Nesse passo supõe-se que a taxa de aprendizado tem efeito fixo, ou seja, é a mesma para todos os indivíduos.
3. Inserção das variáveis explicativas do nível do aluno que podem variar de uma ocasião para outra.
4. Inserção das variáveis explicativas do nível do aluno, invariantes no tempo.
5. Avaliação do efeito aleatório da taxa de aprendizagem. A inclusão desse termo permite que a taxa varie entre os estudantes.
6. Inserção das interações entre as variáveis relativas aos alunos e o tempo.

Neste estudo foram utilizadas apenas características dos alunos que são invariantes no tempo, dessa forma o passo 3 foi suprimido. As variáveis independentes foram centralizadas utilizando as médias gerais correspondentes.

Resultados

Comparação de perfis

A amostra de estudantes selecionada foi dividida em 4 grupos de acordo com o padrão de ausência observado. Esses padrões foram definidos a partir dos registros dos testes de proficiência nos três anos do acompanhamento. Da forma com que a amostra foi coletada, todos os estudantes estavam presentes na avaliação de 2009, logo todos têm o registro dos testes de proficiência e as informações referentes ao questionário contextual naquele ano. Os dados referentes aos demais anos estão incompletos (Tabela 2), ou seja, parte dos estudantes não estava presente nas avaliações realizadas em 2010 e 2011.

Tabela 2
Padrões de ausência de dados.

Padrão	2009	2010	2011	Número de estudantes	%
0	P	P	P	6.447	74,3%
1	P	A	P	450	5,2%
2	P	P	A	652	7,5%
3	P	A	A	1.132	13,0%
Total				8.681	100%

Nota: P = Presente e A = Ausente.

Verifica-se que 74,3% dos estudantes estavam presentes nos três anos do acompanhamento, os demais têm pelo menos uma informação faltante no período 2010-2011. Dada a limitação imposta pelos dados, não foi possível verificar se a ausência relativa aos padrões 1, 2 e 3 foi causada pelo abandono, evasão escolar ou outros motivos[14]. Porém, com a informação disponível, é possível afirmar que o padrão 1 de ausência não foi causado pela evasão escolar uma vez que esses estudantes estavam presentes em 2011. A ausência observada nos padrões 2 e 3 pode ser consequência da

[14] Os dados do SPAECE também foram utilizados no estudo de Shirasu (2014). A autora ressalta a mesma limitação imposta por esses dados.

evasão escolar pois esses estudantes não foram avaliados no último ano, mas também pode ter sido causada pelo abandono e por outros motivos desconhecidos.

Inicialmente foram utilizadas as informações obtidas a partir dos questionários contextuais para comparar os perfis dos estudantes. Na Tabela 3 são apresentadas as distribuições das variáveis onde foram encontradas as maiores diferenças entre os grupos. Outras variáveis foram utilizadas nessa comparação, mas as diferenças encontradas entre os grupos são menores. Esses resultados estão na Tabela A.1 do Apêndice 1.

Tabela 3
Distribuição das variáveis relacionadas aos alunos e suas famílias, por grupo.

Variável	Padrão 0	Padrão 1	Padrão 2	Padrão 3	Total
Sexo					
Feminino	**55,7%**	**54,1%**	48,5%	48,6%	54,1%
Masculino	44,3%	45,9%	**51,5%**	**51,4%**	45,9%
Turno					
Manhã	**43,4%**	33,1%	33,1%	25,4%	39,7%
Tarde	36,1%	35,6%	31,0%	32,2%	35,2%
Noite	20,5%	**31,3%**	**35,9%**	**42,4%**	25,1%
Escolaridade da mãe					
Nunca estudou	5,6%	7,4%	**9,0%**	**9,3%**	6,5%
Não sei.	14,7%	13,0%	**17,5%**	**17,2%**	15,1%
1ª a 4ª do EF	31,8%	31,0%	**34,7%**	31,4%	32,0%
5ª a 8ª do EF	22,8%	22,0%	18,9%	22,5%	22,4%
1ª a 3ª do EM	**18,3%**	**19,8%**	15,2%	14,6%	17,7%
Superior	**6,8%**	**6,7%**	4,7%	5,1%	6,4%
Reprovações					
Nenhuma	**68,6%**	56,9%	44,5%	39,7%	62,4%
Uma	21 ,8%	**27,3%**	**32,2%**	**30,9%**	24,1%
Duas	7,4%	**11,3%**	**17,5%**	**21,2%**	10,1%
Três ou mais	2,2%	4,4%	**5,8%**	**8,2%**	3,4%
Trabalho					
Não	**87,8%**	76,2%	77,3%	71,9%	84,4%
Sim	12,2%	**23,8%**	**22,7%**	**28,1%**	15,6%
Superior					
Não	54,9%	**62,8%**	**71,3%**	**72,5%**	58,8%
Sim	**45,1%**	37,2%	28,7%	27,5%	41,2%

Os valores em negrito da Tabela 3 indicam os percentuais relativos a cada grupo superiores aos observados para todos os estudantes da amostra (última coluna). De forma geral, verifica-se que os estudantes com padrão de ausência 2 e 3 apresentam perfis semelhantes entre si, e distantes do perfil observado para o padrão 0. Os estudantes ausentes apenas no ano de 2010 (padrão 1) apresentam distribuições mais próxima dos estudantes dos padrões 2 e 3, com exceção das variáveis sexo e escolaridade da mãe.

Comparando as distribuições dos padrões 2 e 3 em relação aos demais, pode-se verificar maior percentual de estudantes do sexo masculino, que estudam no período noturno, cujas mães têm baixa escolaridade, com mais reprovações, que trabalham e não pretendem ingressar no ensino superior.

As distribuições das variáveis *reprovações*, *turno*, *trabalho* e *superior* são bastante distintas, principalmente quando são comparados os estudantes dos padrões 0 e 3. Por exemplo, 68,8% dos estudantes do padrão 0 reportaram nenhuma reprovação e esse percentual é igual a 39,7% no padrão 3. Por outro lado, apenas 7,4% dos estudantes com padrão 0 foram reprovados em dois anos letivos e esse percentual é quase três vezes maior para os estudantes do padrão 3 (21,2%).

Na comparação entre os perfis dos estudantes, a *idade* é uma característica importante. A idade média dos alunos presentes em todos os anos da avaliação é 15,6 anos, os estudantes dos outros grupos são mais velhos, com média de 16,5 anos para o grupo com padrão 3. Pela Figura 1, nota-se que a maior frequência relativa é observada na faixa de 15 anos para o primeiro grupo, na faixa de 16 anos para os padrões 1 e 2, e para o último grupo existe maior frequência de estudantes com 17 anos.

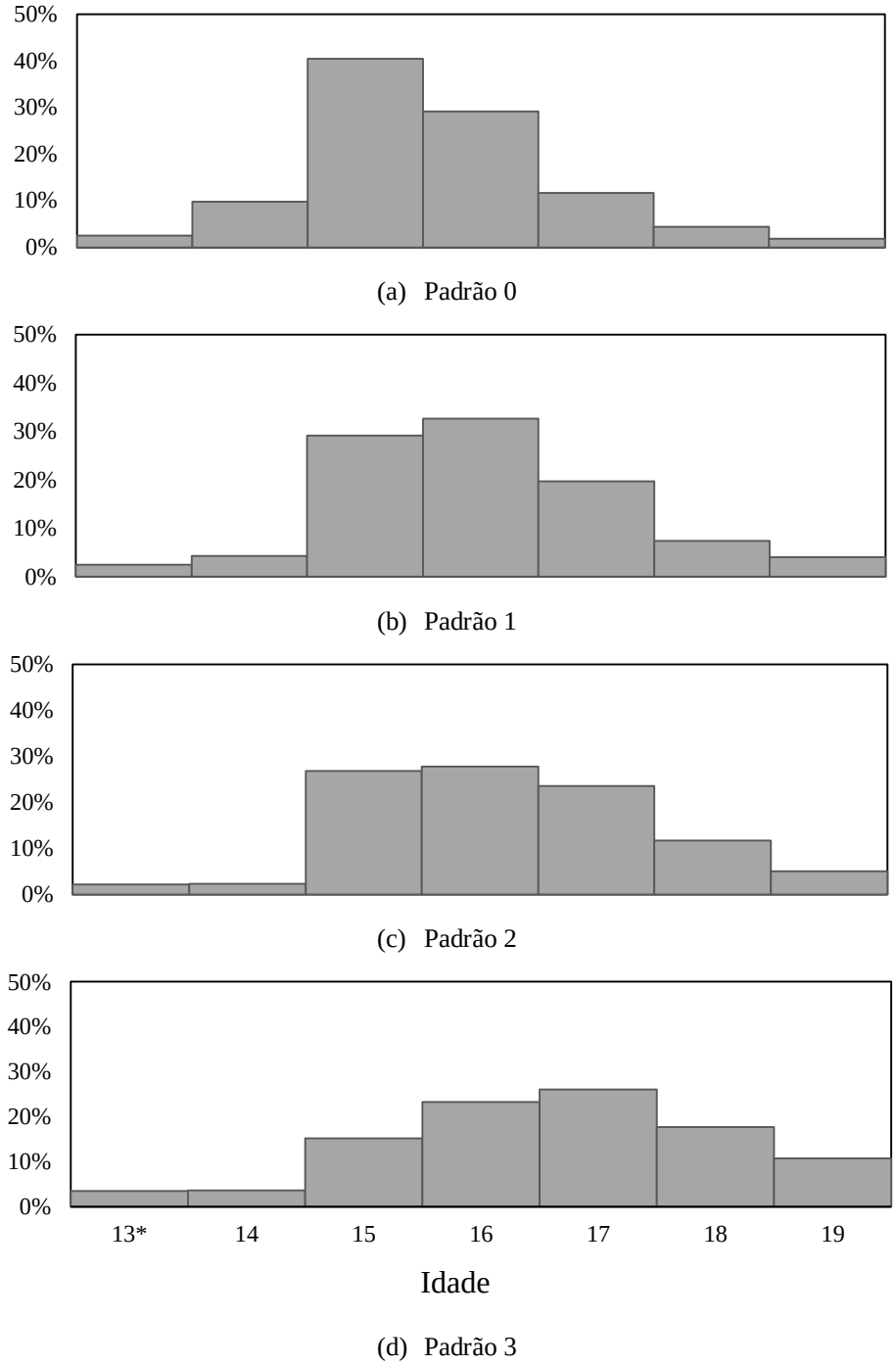

Figura 1
Distribuição da idade dos estudantes, por grupo.
** A primeira faixa contém alunos com 13 anos ou menos.*

Os grupos formados a partir do padrão de ausência também diferem em termos de desempenho escolar (Figura 2). Considerando a proficiência em matemática avaliada em 2009, verifica-se que os estudantes presentes em todo o acompanhamento (padrão 0) tendem a ter melhores desempenhos.

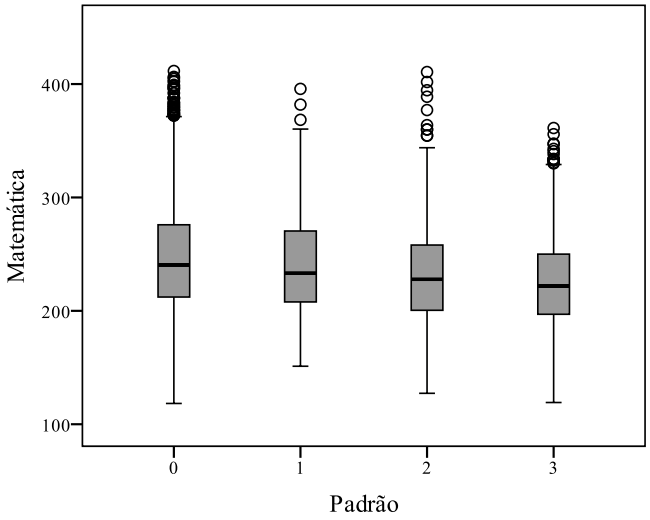

Figura 2
Proficiência em Matemática em 2009, por grupo.

Os resultados apresentados na Tabela 4 confirmam a diferença entre os grupos. Nota-se que os estudantes ausentes em duas ocasiões (padrão 3) e os ausentes somente em 2011 (padrão 2) têm os piores desempenhos médios em matemática. Comparando os padrões 0 e 3, verifica-se que a proficiência média do primeiro grupo é mais de 20 pontos maior (*d* de Cohen igual a 0,48). Pela Figura 2 e Tabela 4, pode-se observar que a variabilidade das notas nas duas disciplinas difere pouco entre os grupos e a maior diferença é observada entre os grupos com padrão 0 e 3.

Tabela 4
Medidas resumo para a proficiência em Matemática em 2009, por grupo.

Variável		Padrão 0	Padrão 1	Padrão 2	Padrão 3	Total
MAT09	Média	246,38	240,73	232,90	225,93	242,41
	Desvio Padrão	46,32	43,59	44,10	39,46	45,76

A evolução do desempenho dos estudantes dos quatro grupos também foi comparada. Pela Figura 3, pode-se verificar que os alunos presentes nas três avaliações realizadas no ensino médio (padrão 0) têm proficiência média maior que os demais e apresentam maior taxa de crescimento no período em relação aos estudantes com padrão 1 e 2 de ausência.

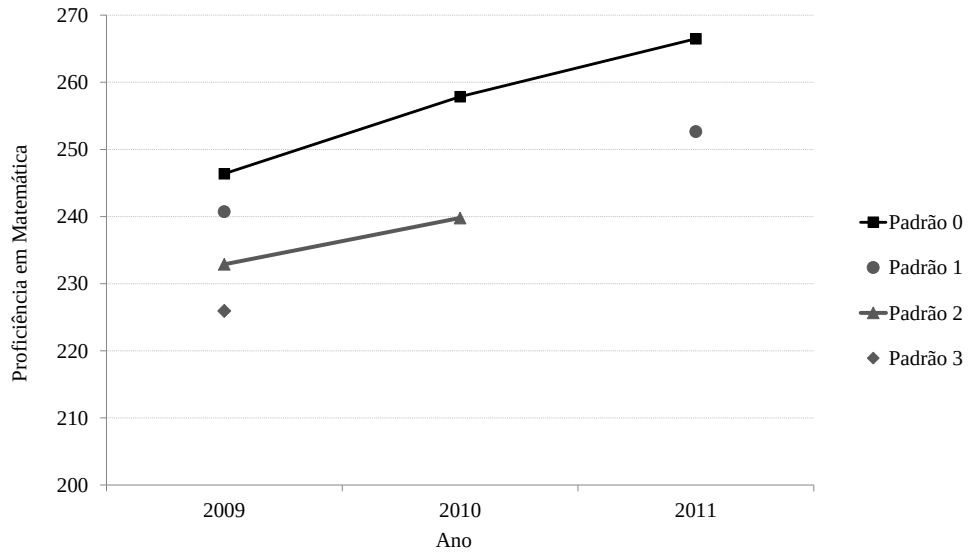

Figura 3
Proficiência média dos estudantes dos diferentes padrões de ausência

Esses resultados mostram que a ausência de dados está relacionada com as variáveis coletadas no estudo. Os grupos diferem tanto em características individuais e familiares quanto no desempenho escolar. Nesse caso, pode-se afirmar que os dados ausentes são não completamente ao acaso (tipo MCAR). No entanto, não é possível avaliar se a ausência de dados é do tipo MAR ou MNAR, pois essa avaliação dependente dos valores não observados (Vinha, 2016).

Neste estudo, assume-se que os dados ausentes são do tipo MNAR, ou seja, que a ausência de informação também está relacionada com os valores que seriam observados para o desempenho. Essa suposição parece adequada tendo em vista os resultados apresentados nessa seção e as conclusões de outros trabalhos onde o desempenho escolar foi identificado como um dos fatores relacionados a evasão e abandono escolar (Soares & colaboradores, 2015).

Modelo de crescimento linear

Inicialmente buscou-se avaliar a parcela da variabilidade dos dados que poderia ser atribuída às escolas. Para isso foi ajustado um modelo considerando três níveis de

agregação: escola, aluno e ocasião. O efeito escola estimado foi 10,5%, o que significa que aproximadamente um décimo da variação do desempenho dos alunos nos anos do acompanhamento pode ser atribuído às escolas. Quando considerada uma proficiência anterior, nesse caso a nota obtida pelos alunos no final do ensino fundamental, o efeito escola estimado é ainda menor representando apenas 6,3% da variação do desempenho. Como o percentual relativo às escolas é baixo, optou-se pelo ajuste de um modelo com apenas dois níveis (aluno e ocasião).

As Tabelas 5 e 6 apresentam os resultados do ajuste do modelo de acordo com os passos propostos por Hox (2010). A escolha do modelo em cada passo baseou-se na comparação dos modelos por meio do teste da diferença de *deviances* e na razão entre o efeito e erro padrão estimados (Raudenbush & Bryk, 2002). Os resultados dos testes são apresentados no Apêndice 1 (Tabela A.2). Por meio da análise dos resíduos verificou-se as suposições do modelo estavam satisfeitas.

A Tabela 5 apresenta os resultados do modelo nulo e o modelo com a inclusão da variável tempo. Pelos resultados do modelo nulo, observou-se uma correlação intraclasse igual a 0,61 (1.462,70/2.382,97), o que significa que a maior parte da variabilidade do desempenho em matemática deve-se à variabilidade entre os indivíduos. No Modelo 1 foi adicionada a variável tempo (o que corresponde ao termo Aprendizado - π_{1i} nas tabelas), considerando apenas seu efeito fixo. Por esse modelo, pode-se afirmar que a proficiência esperada no início do ensino médio é 242,80 pontos e que, em média, a taxa de aprendizado dos estudantes é 8,96 pontos por ano. Nota-se, com a inclusão do tempo, uma redução da variação no primeiro nível e um pequeno aumento da variação entre indivíduos (segundo nível)[15].

[15] Segundo Hox (2010), esse resultado é esperado uma vez que o modelo nulo subestima a variação entre os indivíduos e superestima a variação entre as ocasiões.

Tabela 5
Modelos de crescimento linear (modelos 0 e 1).

	Modelo 0			Modelo 1		
Efeito fixo	Efeito	EP	Razão	Efeito	EP	Razão
Condição inicial (π_{0i})	251,75	0,467	538,98	242,80	0,509	478,10
Aprendizado (π_{1i})				8,96	0,288	31,09
Efeito aleatório						
Nível 1 – ocasião	920,27			840,05		
Nível 2 – aluno	1.462,70			1.489,44		
Total	2.382,97			2.329,49		

No Modelo 2 foram incluídas as variáveis relativas aos estudantes (Tabela 6). Verifica-se que, em relação ao modelo 1, essas variáveis explicam aproximadamente 60% da variação do desempenho entre os estudantes. O efeito estimado para a proficiência anterior é de 0,57 pontos. Esse valor representa o aumento médio na proficiência inicial para cada ponto a mais na nota obtida em 2008. Como a variável *MAT08* tem desvio padrão de aproximadamente 40 pontos, o aumento de um desvio padrão nessa nota resulta em um aumento de 22,8 pontos na proficiência inicial, em média.

Os estudantes que frequentam as aulas no período matutino têm desempenho superior aos demais, em média 9,42 e 4,38 pontos a mais em relação aos que estudam no período noturno e vespertino, respectivamente. Como era esperado, verifica-se o impacto negativo das variáveis relacionadas a reprovação, idade e trabalho, e efeitos positivos da variável *sexo* e *escolaridade_i*. Por fim, mantidas constantes as demais características, os estudantes que pretendem ingressar no ensino superior têm, em média, 8,95 pontos a mais na proficiência em matemática.

Vale ressaltar que as variáveis *etnia* e *itens* não apresentaram efeitos significativos. Provavelmente isso se deve a relação existente entre essas variáveis e outras variáveis preditoras presentes nos modelos.

Tabela 6
Modelos de crescimento linear (modelos 2, 3 e 4).

	Modelo 2			Modelo 3			Modelo 4		
Efeito fixo	Efeito	Erro Padrão	Razão	Efeito	Erro Padrão	Razão	Efeito	Erro Padrão	Razão
Condição inicial (π_{0i})									
Intercepto	242,80	0,421	577,36	242,80	0,390	621,98	242,80	0,390	622,71
MAT08	0,57	0,009	60,56	0,57	0,009	64,84	0,57	0,009	64,84
Sexo	8,07	0,778	10,37	7,97	0,739	10,79	7,97	0,739	10,79
Escolaridade_i	2,05	0,922	2,23	2,00	0,869	2,30	2,00	0,869	2,30
Idade_i	-7,59	1,096	-6,92	-7,00	1,030	-6,80	-5,43	1,081	-5,02
Tarde	-4,38	0,781	-5,61	-4,39	0,760	-5,77	-4,39	0,760	-5,77
Noite	-9,42	1,018	-9,25	-9,01	0,970	-9,30	-7,86	1,031	-7,62
Reprovação_i	-6,83	0,943	-7,25	-6,37	0,895	-7,12	-5,15	0,949	-5,42
Trabalho	-3,36	1,152	-2,92	-3,35	1,094	-3,07	-3,35	1,094	-3,07
Superior	8,95	0,778	11,51	8,83	0,742	11,90	8,83	0,742	11,90
Aprendizado (π_{1i})									
Intercepto	8,96	0,288	31,09	8,96	0,292	30,68	8,96	0,291	30,77
Idade_i							-2,15	0,802	-2,68
Noite							-1,58	0,669	-2,36
Reprovação_i							-1,68	0,713	-2,35
Efeito aleatório									
Nível 1 – ocasião		840,05			800,13			800,12	
Nível 2 – aluno		581,93			399,08			397,31	
Aprendizado					39,92			36,61	
Total		1421,98			1239,13			1234,04	

O componente aleatório da taxa de aprendizado foi introduzido no Modelo 3 (Tabela 6). Observa-se que a variância estimada para esse componente é igual a 39,92, o que corresponde a um desvio padrão de 6,32, aproximadamente 70% do efeito fixo estimado (8,96). Logo, pode-se afirmar que a taxa de aprendizagem estimada varia consideravelmente entre os estudantes.

Na Tabela 6 também são apresentados os resultados do modelo com a inclusão dos efeitos de interação entre a ocasião e algumas variáveis dos estudantes (Modelo 4). Observa-se que idade, o número de reprovações e turno da noite reduzem significativamente a taxa de aprendizado. Por exemplo, mantidas constantes as demais variáveis, os estudantes mais velhos têm taxa de crescimento média igual a 6,81 pontos por ano letivo (8,96 – 2,15 = 6,81).

Discussão

O objetivo principal deste trabalho foi identificar fatores relacionados ao desempenho escolar dos estudantes do ensino médio utilizando os dados de uma avaliação educacional em larga escala. Estudos semelhantes foram realizados no Brasil (Karino & Laros, 2016; Vinha & Laros, 2016) e os resultados aqui apresentados são compatíveis com os observados na literatura. Porém, vale ressaltar que este estudo apresentou a análise de dados longitudinais, ainda raros na avaliação em larga escala no Brasil, e discutiu o tratamento e a análise de dados com observações ausentes neste contexto.

Os dados utilizados apresentaram muitos valores ausentes. Por isso, visando avaliar o tipo de ausência observada, foi realizada uma comparação dos perfis dos estudantes agrupados de acordo com padrão de ausência. Verificou-se que a ausência está relacionada com variáveis presentes no estudo como sexo, escolaridade dos pais, trabalho, reprovações e o próprio desempenho escolar. Esse fato foi considerado na escolha do método de análise utilizado.

Dada a limitação imposta pelos dados disponibilizados, não foi possível avaliar com clareza os fatores relacionados com evasão ou abandono escolar nas escolas públicas do Ensino Médio Cearense. Contudo, supõe-se que uma parcela considerável dos estudantes presentes apenas na avaliação de 2009 tenha se evadido da escola, assim o perfil desse grupo está próximo do perfil de estudantes evadidos. Os resultados observados para esse grupo de estudantes são semelhantes aos observados em estudos sobre evasão e abandono escolar (De Leon & Menezes-Filho, 2002; Gonçalves, 2008; Shirasu, 2014; Soares & cols, 2015).

O presente estudo se destaca ao utilizar os dados de um levantamento longitudinal. Nas avaliações longitudinais, a medida da proficiência anterior é usada para filtrar o

aprendizado prévio do estudante, tornando possível avaliar o conhecimento adquirido no período de interesse. No entanto, os resultados observados são semelhantes aos observados por autores que avaliaram o desempenho dos estudantes no final do ensino médio com dados transversais.

As variáveis relacionadas ao nível socioeconômico, sexo, etnia e escolaridade da mãe são frequentemente utilizadas neste tipo de estudo, e em alguns casos são usadas como variáveis de controle (Vinha e Laros, 2016). A relação significativa entre a proficiência e a escolaridade da mãe do estudante também foi identificada neste estudo. Ainda, verificou-se que os estudantes do sexo masculino têm melhor desempenho em matemática, em média. No entanto, os resultados não indicaram um impacto significativo das variáveis *itens* (usada como um indicador do nível socioeconômico) e *etnia*. Vale lembrar que a proficiência anterior dos estudantes foi utilizada no ajuste dos modelos, como uma variável de controle, o que muito provavelmente anulou o efeito de outras características. Por exemplo, estudantes autodeclarados brancos ou amarelos tinham desempenho médio superior no final do ensino fundamental, logo no ajuste do modelo de crescimento linear a variável *etnia* não apresenta efeito significativo.

O efeito escola estimado no presente estudo é menor que o observado por outros autores. Observou-se um efeito escola bruto igual a 10,5% e, quando considerada a proficiência anterior, esse efeito passou para 6,3%. Andrade e Laros (2007) estimaram um efeito escola bruto de 46%, e após a introdução de variáveis de controle, o efeito estimado foi de 17%. Ainda com os dados do SAEB, Gonçalves e França (2008) estimaram um efeito escola igual a 42,3%. Essas diferenças podem ser explicadas pela maior homogeneidade das escolas avaliadas no SPAECE. Na edição de 2001 e 2003 do SAEB foram selecionadas escolas públicas e privadas de todo o país, enquanto que no SPAECE apenas as escolas da rede pública do estado do Ceará estão envolvidas. Ainda,

a diferença observada no efeito escola também pode ser atribuída ao tipo de levantamento e ao período de realização da avaliação do SPAECE.

A exemplo de outros trabalhos, o presente estudo verificou o efeito de variáveis relacionadas com o histórico escolar dos estudantes no desempenho. Verificou-se que os alunos que reportaram episódios de reprovação e atrasados (com 17 anos ou mais no primeiro ano do estudo) têm desempenhos médios inferiores. Com base nas avaliações do 3º ano do ensino médio do SAEB, Andrade e Laros (2007) e Gonçalves e França (2008) identificaram o impacto negativo do atraso escolar e do número de reprovações no desempenho dos estudantes; e Laros, Marciano e Andrade (2012) também verificaram o impacto negativo do atraso, abandono escolar e repetência.

Vale ressaltar que a análise dos dados de um acompanhamento longitudinal possibilitou a identificação de variáveis que influenciam a taxa de aprendizado dos estudantes durante o ensino médio. Verificou-se que os estudantes mais velhos, com reprovações no histórico escolar e que estudam à noite têm menor taxa de aprendizado.

Por fim, a partir dos resultados apresentados, este estudo confirma a importância da reprovação e do atraso escolar na evolução dos estudantes do ensino médio. A idade e o número de reprovações são fortes preditoras do baixo desempenho escolar (tanto desempenho inicial, quanto na taxa de aprendizado), além de estarem associadas a maiores taxas de ausência.

Considerações finais

Este estudo apresenta algumas limitações relacionadas aos dados utilizados. Primeiro, os bancos de dados dos diferentes anos não têm uma identificação comum para os estudantes, nesse caso a interligação foi feita com base nos nomes dos alunos através de um procedimento passível de erros (Vinha, 2016). Segundo, com esses dados não foi possível verificar se a ausência dos estudantes na avaliação se deve a evasão ou

ao abandono escolar. Ainda, o questionário contextual utilizado em 2011 difere dos questionários dos outros anos, dessa forma não foi possível incluir variáveis que avaliavam, por exemplo, o clima em sala de aula ou a relação entre os estudantes e os professores.

Resultados observados e limitações impostas pelo banco de dados sinalizam algumas oportunidades de pesquisas futuras. Por exemplo, o método utilizado baseia-se na suposição de que os dados ausentes são do tipo MNAR e o procedimento utilizado supõe ainda que a evolução dos estudantes ausentes seria semelhante à evolução dos estudantes com reprovações no ensino médio (Vinha, 2016). Portanto, em pesquisas futuras, sugere-se a utilização de outras abordagens de análise para esses dados e a comparação com os resultados aqui apresentados. Outra sugestão para trabalhos futuros está relacionada às variáveis independentes. No presente estudo foram utilizadas apenas variáveis coletadas no primeiro ano do acompanhamento, e essas variáveis foram consideradas invariantes no tempo. Variáveis independentes que podem variar de um período para outro, como horas de estudo ou percentual de faltas, e fatores relacionados à vida escolar poderiam ser utilizados.

Espera-se que o presente estudo possa contribuir para o avanço do ensino médio no país, fornecendo informações que podem ser usadas no delineamento de programas e ações de melhoria. Espera-se também contribuir com a discussão sobre a ausência de informação e a utilização de novas metodologias para a análise dos dados das avaliações educacionais.

Referências

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Andrade, J. M., & Laros, J. A. (2007). Fatores associados ao desempenho escolar: Um estudo multinível com os dados do SAEB/2001. *Psicologia: Teoria e Pesquisa, 23,* 33-42.

Ceará (2011). Secretaria da Educação. SPAECE – 2011. *Boletim Pedagógico Matemática - Ensino Médio, 3, 1-22*.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 20*(10), 1-20.

De Leon, F. L. L., & Menezes-Filho, N. A. (2002). Reprovação, avanço e evasão escolar no Brasil. *Pesquisa e Planejamento Econômico, 32*(3), 417-452.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods, 16*(1), 1-16.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2009). *Longitudinal data analysis.* Boca Raton: Chapman & Hall.

Franco, C. (2001). O SAEB: Potencialidades, problemas e desafios. *Revista Brasileira de Educação, 17,* 127-132.

Franco, C., Ortigão, I., Albernaz, A., Bonamino, A., Aguiar, G., Alves, F., & Sátyro, N. (2007). Qualidade e equidade em educação: Reconsiderando o significado de "fatores intra-escolares". *Ensaio: Avaliação e Políticas Públicas em Educação, 15*(55), 277-298.

Goldstein, H. (2010). *Multilevel statistical models* (fourth edition). London: Wiley.

Gonçalves, F. O., & França, M. T. A. (2008). Transmissão intergeracional de desigualdade e qualidade educacional: Avaliando o sistema educacional brasileiro a partir do SAEB 2003. *Ensaio: Avaliação e Políticas Públicas em Educação, 16*(61), 639-662.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549-576.

Hippolyto, L. Q. (2013). *Avaliação dos resultados do 3º ano do ensino médio em Matemática no Ceará e sua repercussão na prática pedagógica dos professores: Um estudo descritivo a partir dos testes do SPAECE nos anos 2008, 2009 e 2010.* Dissertação de Mestrado, Universidade Federal do Ceará, Fortaleza.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (second edition). New York: Routledge.

Instituto de Pesquisa e Estratégia Econômica do Ceará (2014). *Desempenho do Ceará no SAEB: Análise comparativa com os estados brasileiros - 2005 e 2011*. Fortaleza: Instituto de Pesquisa e Estratégia Econômica do Ceará.

Karino, C. A., & Laros, J. A. (2016). Estudos brasileiros sobre eficácia escolar: Uma revisão de literatura. Manuscrito submetido para publicação.

Krawczyk, N. (2011). Reflexão sobre alguns desafios do ensino médio no Brasil hoje. *Cadernos de Pesquisa, 41*(144), 752-769.

Laros, J. A., Marciano, J. L. P., & Andrade, J. M. (2012). Fatores associados ao desempenho escolar em Português: Um estudo multinível por regiões. *Ensaio: Avaliação e Políticas Públicas em Educação, 20,* 623-646.

Lee, V. L. (2008). Utilização de modelos lineares hierárquicos lineares para estudar contextos sociais: O caso dos efeitos da escola. In N. Brooke & J. F. Soares (Eds.), *Pesquisa em eficácia escolar: Origem e trajetórias* (pp. 273-296). Belo Horizonte: Editora UFMG.

Neri, M. (2009). *Motivos da evasão escolar*. Brasília: Fundação Getúlio Vargas.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525-556.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (second edition). Newbury Park: Sage.

Ribeiro, L. V. F., Júnior, F. T. & Lourenço, J. L. (2013). Avaliação e Currículo: um diálogo necessário. *ANAIS - VII Reunião da ABAVE, 1*, 479-496.

Rousseau, M., Simon, M., Bertrand, R., & Hachey, K. (2012). Reporting missing data: A study of selected articles published from 2003-2007. *Quality & Quantity, 46*(5), 1393-1406.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Shirasu, M. R. (2014). *Determinantes da evasão e repetência escolar no Ceará*. Dissertação de Mestrado. Universidade Federal do Ceará, Fortaleza.

Soares, J. F., & Candian, J. (2007). O efeito da escola básica brasileira: As evidências do PISA e do SAEB. *Revista Contemporânea de Educação, 2*(4), 45-64.

Soares, T. M., Fernandes, N. S., Nóbrega, M. C., & Nicolella, A. C. (2015). Fatores associados ao abandono escolar no ensino médio público de Minas Gerais. *Educação e Pesquisa, 41*(3), 757-772.

Todos pela Educação (2015). *De olho nas metas 2013-14: Sexto relatório de monitoramento das 5 metas do Todos pela Educação*. São Paulo: Moderna.

Vinha, L. G. A., Karino, C. A., & Laros, J. A. (no prelo). Fatores associados ao desempenho em matemática na educação Brasileira: Um estudo multinível. *Psico-USF,* Manuscrito aceito para publicação.

Vinha, L. G. A. (2016). *Tratamento de dados ausentes em uma avaliação educacional com dados longitudinais*. Manuscrito em preparação.

Vinha, L. G. A., & Laros, J. A. (2016). *Avaliações educacionais no Brasil, Chile e Argentina: uma revisão de literatura*. Manuscrito submetido para publicação.

Volpi, M., Silva, M. S., & Ribeiro, J. (2014). *10 desafios do ensino médio no Brasil: Para garantir o direito de aprender de adolescentes de 15 a 17 anos.* Brasília: UNICEF.

Soares, J. F., Alves, M. T. G., & Oliveira, R. M. (2001). O efeito de 248 escolas de nível médio no vestibular da UFMG nos anos de 1998, 1999 e 2000. *Estudos em Avaliação Educacional, 24*, 69-117.

Apêndice 1

Tabela A.1
Distribuição de variáveis relacionadas aos alunos, por grupo.

Variável	Padrão 0	Padrão 1	Padrão 2	Padrão 3	Total
Computador e Internet					
Sim, com Internet	**10,3%**	**10,7%**	9,5%	8,7%	10,0%
Sim, sem Internet	**6,0%**	**6,0%**	3,7%	4,8%	5,7%
Não	82,4%	82,9%	**84,7%**	**84,7%**	82,9%
Número de pessoas					
3 ou menos	17,1%	18,3%	**19,0%**	**22,4%**	18,0%
4 pessoas	**26,2%**	24,5%	22,1%	21,0%	25,2%
5 pessoas	**25,1%**	23,6%	21,0%	22,1%	24,3%
6 pessoas	15,8%	15,1%	17,5%	14,5%	15,8%
7 ou mais	15,7%	**18,5%**	**20,4%**	**20,0%**	16,7%
Livros					
1 a 10 livros	35,6%	35,0%	34,4%	**37,0%**	35,7%
11 a 20 livros	23,0%	**24,8%**	**26,7%**	**24,0%**	23,5%
21 e 50 livros	**18,3%**	17,6%	15,0%	15,8%	17,7%
Mais de 50 livros	**13,5%**	**13,4%**	**13,6%**	12,2%	13,3%
Nenhum	9,6%	9,2%	**10,3%**	**10,9%**	9,8%
Disciplina que mais gosta					
Língua Portuguesa.	23,5%	**26,6%**	**25,6%**	**27,4%**	24,4%
Matemática.	**19,7%**	14,9%	16,4%	15,9%	18,7%
Educação Física.	21,0%	21,0%	**26,1%**	**25,1%**	21,9%
Outra.	**35,8%**	**37,5%**	32,0%	31,5%	35,0%

Tabela A1.2
Resultados dos testes para escolha dos modelos.

	Modelo 0	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Deviance	271246,77	269.637,23	263.280,29	262.930,42	262859,10
Nº de parâmetros	3	4	13	14	17
Diferença		1.609,54	6.356,94	349,87	71,32
Estatística χ^2		1.609,54	706,33	349,87	23,77

Considerações finais

O presente estudo teve como objetivo discutir e aplicar metodologias para tratamento de valores ausentes com ênfase na análise dos dados de avaliações educacionais. Essa discussão teve como contexto o estudo de fatores associados ao desempenho escolar a partir dos dados da avaliação do ensino médio realizada no estado do Ceará (SPAECE). A seguir são apresentados alguns comentários, recomendações e as limitações do estudo.

Os métodos tradicionais de tratamento de dados ausentes ainda são muito utilizados pelos pesquisadores, apesar das críticas frequentes. No presente estudo, a imputação pela média e o procedimento *listwise deletion* foram comparados com métodos sofisticados. A imputação pela média apresentou desempenho insatisfatório em todos os padrões de ausência simulados (Manuscrito 1). O uso desse procedimento gera estimativas viesadas para os parâmetros do modelo, mesmo quando a ausência é completamente ao acaso. Esse resultado corrobora a afirmação de Enders (2010) de que a imputação pela média é possivelmente o pior tratamento para dados ausentes.

Pode-se observar que o procedimento *listwise deletion* apresentou desempenho satisfatório no ajuste do modelo de regressão quando os dados ausentes simulados são do tipo MCAR ou MAR. Esse procedimento tem desempenho inferior nos cenários onde a ausência simulada é do tipo MNAR, mas isso acontece somente quando são utilizadas variáveis auxiliares na estimação por máxima verossimilhança e na imputação múltipla. No segundo manuscrito, o *listwise deletion* é comparado com outros métodos na análise dos dados do SPAECE, onde a ausência observada é considerada não ao acaso (MNAR). Nessa comparação, as diferenças entre os resultados do *listwise deletion* e das outras metodologias são mais evidentes. Deve-se considerar o uso desse procedimento quando a ausência de dados é completamente ao acaso ou em algumas

situações onde a ausência é do tipo MAR e as variáveis relacionadas com a ausência estão presentes na modelagem.

O estudo apresentado no primeiro manuscrito também pode ser usado para avaliar o impacto do percentual de valores ausentes nas análises. Excluindo os resultados relativos a imputação pela média, pode-se observar que as estimativas obtidas para os parâmetros do modelo de regressão são pouco afetadas quando 10% dos indivíduos da amostra apresentam valores ausentes. No entanto, esse resultado não pode ser usado como parâmetro para outras situações, pois depende da técnica utilizada e do padrão de ausência. Segundo Hair, Black, Babin e Anderson (2010), em geral, a presença dos dados ausentes pode ser ignorada quando ocorre em até 10% das observações, outros autores são mais conservadores e adotam como limite apenas 5% de observações ausentes (Tabachnick e Fidell, 2006).

A ausência de dados na avaliação dos estudantes do ensino médio do Estado do Ceará não é completamente ao acaso. Verificou-se que essa ausência estava relacionada com o perfil e histórico escolar do aluno, o que pode ser explicado pela parcela de estudantes que abandonaram ou se evadiram da escola (manuscrito 3). Porém, assim como em outras situações, não é possível avaliar se os dados ausentes observados no estudo podem ser considerados do tipo MAR ou MNAR. Em outras palavras, a taxa de ausência está relacionada com as características dos alunos e com o desempenho anterior, mas não é possível afirmar se existe ainda uma relação residual entre a ausência e o desempenho que seria avaliado caso o aluno estivesse presente naquele ano. Esse fato motivou a comparação dos resultados obtidos a partir da aplicação da imputação múltipla, que tem como suposição a ausência do tipo MAR, e de uma nova metodologia desenvolvida para a análise de dados MNAR (Manuscrito 2).

O método proposto no segundo manuscrito baseia-se nos modelos de misturas de padrões. Uma suposição adicional se fez necessária para identificação do modelo, nesse caso se supôs que a evolução dos indivíduos ausentes é semelhante à evolução dos indivíduos com reprovações no ensino médio. Essa suposição fundamenta-se em alguns estudos citados neste documento que mostram a relação existente entre desempenho, reprovação, abandono e evasão escolar. A utilização desse novo método mostrou resultados semelhantes aos obtidos pela imputação múltipla, ambos estimam menor taxa de aprendizado dos estudantes e maior impacto das variáveis independentes em relação ao procedimento *listwise deletion*. Uma vez que a diferença entre os métodos está relacionada a suposições não testáveis, não é possível afirmar qual resultado pode ser considerado mais correto.

A experiência adquirida com a análise dos dados do SPAECE demostrou a importância da avaliação do padrão de ausência observado no levantamento. Os dados faltantes podem trazer informações importantes acerca do fenômeno sob investigação e do procedimento utilizado para coleta dos dados. Essa avaliação é fundamental para identificar o mecanismo gerador da ausência e as variáveis que podem ser usadas como auxiliares no tratamento dos dados. Mesmo que o padrão de ausência observado no estudo não seja considerado nas análises, recomenda-se que essa informação seja mencionada no relato da pesquisa, visando explicitar os possíveis impactos nos resultados observados.

A utilização de métodos mais sofisticados para o tratamento de dados ausentes depende da disponibilidade de pacotes estatísticos. No presente estudo, a estimação por máxima verossimilhança e o procedimento de imputação múltipla foram implementados através das funções do software estatístico SAS. No entanto, esses procedimentos podem ser encontrados outros pacotes comerciais como o MPlus e no software livre R.

A oportunidade de analisar os dados de uma avaliação longitudinal motivou a utilização dos dados do SPAECE. No entanto, é importante registrar a limitação imposta pelos dados disponibilizados para a elaboração desta tese. A inexistência de uma identificação única dos estudantes durante o período avaliado não permitiu a junção automática das informações contidas nos arquivos. O procedimento adotado para interligação dos arquivos impôs a redução do número de estudantes na amostra, os dados referentes a 8% dos estudantes foram utilizados. Além disso, esse procedimento é passível de erros por usar os nomes dos alunos como identificador. Sugere-se, portanto, que o planejamento dessas avaliações considere o uso de identificadores numéricos para os estudantes, e que esses identificadores sejam os mesmos em todos os momentos de coleta de dados, de tal forma que a junção dos arquivos possa ser feita automaticamente e sem erros.

Espera-se que este trabalho incentive outros pesquisadores a avaliar os problemas relacionados à ausência de informação e utilizar metodologias adequadas para tratamento de dados ausentes. Assim, espera-se ter contribuído com as discussões acerca da utilização dos dados das avaliações educacionais.

Referências

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Englewood Cliffs: Prentice Hall.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (fifth edition). Boston: Pearson.